PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-179883
(43)Date of publication of application : 11.07.1997

(51)Int.CI. G06F 17/30
G06F 12/00
G06F 12/00

(21)Application number : 08-044660
(22)Date of filing : 01.03.1996

(71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(72)Inventor : FUKUDA TSUYOSHI
MORIMOTO YASUHIKO
MORISHITA SHINICHI
TOKUYAMA TAKESHI

(30)Priority
Priority number : 07278690 Priority date : 26.10.1995 Priority country : JP

(54) METHOD AND DEVICE FOR DERIVING CONNECTION RULE BETWEEN DATA

(57)Abstract:
PROBLEM TO BE SOLVED: To make it possible to find the correlation between data which have a two-term numeral attribute and a true/false attribute.
SOLUTION: A plane is constituted with two numeral attributes first and divided into meshes, and data in the meshes (packet) and data which have true attributes are counted. This plane can be grasped as the plane image that the number of data corresponds to the gray level and the number of data having true attributes corresponds to the saturation. Then a permissible image which is an area that is convex to one axis of the plane is cut under specific conditions and a part where the correlation of data is strong is found. Then when the area as the cut permissible area meets conditions of a support maximization rule, etc., the area is shown to a user. Further, necessary attributes of data included in the area are extracted from a data base at need.

LEGAL STATUS

[Date of request for examination]	13.05.1998
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3118181
[Date of registration]	06.10.2000
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's	

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－179883

(43)公開日　平成９年(1997)７月11日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	17/30			Ｇ０６Ｆ	15/403	３４０Ｚ
	12/00	５０５			12/00	５０５
		５２０				５２０Ａ

審査請求　未請求　請求項の数23　ＯＬ　（全 26 頁）

(21)出願番号　特願平8－44660

(22)出願日　平成８年(1996)３月１日

(31)優先権主張番号　特願平7－278690

(32)優先日　平７(1995)10月26日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

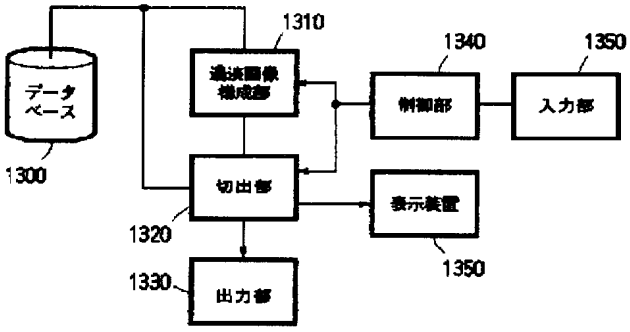
(72)発明者　福田　剛志
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(74)代理人　弁理士　合田　潔　（外２名）

最終頁に続く

(54)【発明の名称】　データ間結合ルール導出方法及び装置

(57)【要約】
【課題】２項の数値属性と真偽をとる属性を有するデータ間の相関を見い出すための手法を提案すること。
【解決手段】（１）２つの数値属性により平面を構成し、この平面をメッシュ分割し、各メッシュ（バケットともいう）内のデータ数及び真偽をとる属性が真となったデータの数をカウントする。このような平面は、各メッシュをピクセルとした場合、データ数が濃淡度、真偽をとる属性が真となるデータの数が彩度に該当するような、平面画像として捉えることができる。（２）所定の条件θに従い、平面の１つの軸に凸な領域である許容イメージを切り出し、データの相関の強い部分を見い出す。（３）切り出した許容イメージとなる領域が、サポート最大化ルール等の条件を満たしていれば、その領域をユーザに提示する。また、データベースからその領域に含まれるデータの必要な属性を引き出すことも、必要に応じて行う。



1

【特許請求の範囲】

【請求項1】２種類の数値属性と、１種類の真偽をとる属性を含むデータを有するデータベースにおいて、
データ間の結合ルールを導き出す方法であって、
前記２種類の数値属性に対応する２つの軸を有し且つN×M個のバケットに分割されている平面の各バケットに対応して、当該バケット（座標(i, j)）に属するデータの数u(i, j)及び前記真偽をとる属性が真であるデータの数v(i, j)を記憶する平面構成ステップと、
条件θを入力するステップと、

【数1】

$$\sum_{(i,j)\in S} g(i,j) = \sum_{(i,j)\in S} (v(i,j) - \theta u(i,j))$$

を最大にするような前記バケットの領域Sを前記平面から切り出す領域切出ステップと、
切り出された前記領域S内に含まれるデータを出力するステップとを含むデータ間結合ルール導出方法。

【請求項2】入力された前記条件θとは異なる第２の条件θ_2を入力するステップと、

【数2】

$$\sum_{(i,j)\in S_2} g(i,j) = \sum_{(i,j)\in S_2} (v(i,j) - \theta_2 u(i,j))$$

を最大にするような前記バケットの第２の領域S_2を前記平面から切り出すステップと、

【数3】

$$\theta 3 = \frac{V(S_2) - V(S)}{U(S_2) - U(S)}$$

（前記領域S_2に含まれ且つ前記真偽をとる属性が真であるデータの数をV（S_2）、前記領域Sに含まれ且つ前記真偽をとる属性が真であるデータの数をV（S）、前記領域S_2に含まれるデータ数をU（S_2）、前記領域Sに含まれるデータ数をU（S）とする。）を第３の条件として、

【数4】

$$\sum_{(i,j)\in S_3} = \sum_{(i,j)\in S_3} (v(i,j) - \theta_3 u(i,j))$$

を最大にするような前記バケットの第３の領域S_3を前記平面から切り出すステップとをさらに含む請求項１記載のデータ間結合ルール導出方法。

【請求項3】前記切り出された領域S内の各バケットのv(i, j)／u(i, j)が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性が真であるデータ数の割合に等しくなるようv(i, j)を変更するステップと、
当該変更されたv(i, j)を用いて、入力された条件θ_4に従い、

【数5】

2

$$\sum_{(i,j)\in S_4} g(i,j) = \sum_{(i,j)\in S_4} (v(i,j) - \theta_4 u(i,j))$$

を最大にするような前記バケットの第４の領域S_4を切り出すステップとをさらに含む請求項１記載のデータ間結合ルール導出方法。

【請求項4】前記平面構成ステップが、
複数の前記データから、X個のデータをランダムサンプリングするステップと、
サンプリングされたデータを各前記数値属性についてソートし、X・i／N（i＝1，2，・・N）番目に該当する数値及びX・n／M（n＝1，2，・・M）番目に該当する数値を記憶するステップと、
記憶された前記数値を基準にして、前記複数のデータをN×M個の前記バケットに入れるステップとを含む請求項１乃至３記載のデータ間結合ルール導出方法。

【請求項5】

【数6】

$$g(i,j) = v(i,j) - \theta u(i,j)$$

を評価値とし、
前記領域切出ステップが、
前記平面の各列において、少なくとも１のバケットを含み且つ前記評価値が最大となるバケットの範囲を求める列内範囲導出ステップと、
前記平面内の任意のバケット（座標(m, t)）の前列にあるもう１つのバケット（座標(m-1, l)）を含み且つ当該前列（第１列）までで最大の評価値を有する領域と、前記列内範囲導出ステップにより導出される、前記任意のバケット（座標(m, t)）及び前記もう１つのバケットと同一行であって前記任意のバケットと同一列にあるバケット（座標(m, l)）を含む評価値が最大となる範囲とを加えた領域を加算領域とする時、各前記任意のバケットに対し、それ自身及び当該任意のバケットを含む前記加算領域全体として評価値を最大化する前記もう１つのバケットを検出し、当該加算領域の評価値と共に記憶する検出記憶ステップと、
前記平面全体で前記加算領域の評価値が最大となる前記任意のバケットを検出し、前記検出記憶ステップで検出された前記もう１つのバケットを用いて、領域Sを導き出すステップとを含む請求項１記載のデータ間結合ルール導出方法。

【請求項6】切り出されるべき領域に含まれる最低限のデータ数である最小サポート数Uminを入力するステップと、
前記切り出された領域Sに含まれるデータ数U（S）と前記最小サポート数Uminと比較するステップと、
前記比較の結果、Umin≒U（S）であれば、当該領域Sを切り出されるべき領域として出力するステップと、
前記比較の結果、Umim＞U（S）又はUmim＜U（S）の場合には、新たな条件θ_5にて、前記領域切出ステップを実施するステップとをさらに含む請求項１記載のデ

ータ間結合ルール導出方法。

【請求項7】切り出されるべき領域における前記真偽をとる属性が真であるデータの数の割合minconfを入力するステップと、

前記切り出された領域Sが、minconf≒V（S）／U（S）（U（S）は前記領域Sに含まれるデータ数、V（S）は前記領域Sに含まれ且つ前記真偽をとる属性が真であるデータの数）である場合には、当該領域Sを出＊力するステップと、

minconf＜V（S）／U（S）又はminconf＞V（S）／U（S）である場合には、新たな条件θ₁にて前記領域切出ステップを実施するステップとをさらに含む請求項1記載のデータ間結合ルール導出方法。

【請求項8】切り出された前記領域Sに対し、

【数7】

$$f(U(S),V(S)) = -V(S)\log\frac{V(S)}{U(S)} - (U(S)-V(S))\log\frac{U(S)-V(S)}{U(S)} - (Vsum-V(S))\log\frac{Vsum-V(S)}{Usum-U(S)} - (Usum-Vsum-U(S)+V(S))\log\frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}$$

（Usumは前記平面全体のデータ数、Vsumは前記平面全体に含まれる前記真偽をとる属性が真のデータの数）を計算し、その値を前記領域Sに対応して記憶するエントロピ計算ステップと、

条件θを変更して前記領域切出ステップと前記エントロピ計算ステップを実行するステップと、

※f（U（S），V（S））を最大化する領域Sを出力するステップとをさらに含む請求項1記載のデータ間結合ルール導出方法。

【請求項9】切り出された前記領域Sに対し、

【数8】

$$f(U(S),V(S)) = U(S)\left(\frac{V(S)}{U(S)} - \frac{Vsum}{Usum}\right)^2 + (Usum-U(S))\left(\frac{Vsum-V(S)}{Usum-U(S)} - \frac{Vsum}{Usum}\right)^2$$

（Usumは前記平面全体のデータ数、Vsumは前記平面全体に含まれる前記真偽をとる属性が真のデータの数）を計算し、その値を前記領域に対応して記憶するインタクラスバリアンス計算ステップと、

条件θを変更して前記領域切出ステップと前記インタクラスバリアンス計算ステップとを実行するステップと、

f（U（S），V（S））を最大化する領域Sを出力するステップとをさらに含む請求項1記載のデータ間結合ルール導出方法。

【請求項10】2種類の数値属性と、1種類の真偽をとる属性を含むデータを有するデータベースにおいて、

データ間の結合ルールを導き出す装置であって、

前記2種類の数値属性に対応する2つの軸を有し且つN×M個のバケットに分割されている平面の各バケットに対応して、当該バケット（座標(i，j)）に属するデータの数u(i，j)及び前記真偽をとる属性が真であるデータの数v(i，j)を記憶する平面構成手段と、

条件θを入力する入力手段と、

【数9】

$$\sum_{(i,j)\in S} g(i,j) = \sum_{(i,j)\in S} (v(i,j)-\theta u(i,j))$$

を最大にするような前記バケットの領域Sを前記平面から切り出す領域切出手段と、

切り出された前記領域S内に含まれるデータを出力する手段と、

を有するデータ間結合ルール導出装置。

【請求項11】前記入力手段により、前記条件θとは異なる第2の条件θ₂を入力し、前記領域切出手段により、前記第2の条件θ₂に対応する第2の領域S₂を前記平面から切り出した場合に、

【数10】

$$\theta 3 = \frac{V(S_2)-V(S)}{U(S_2)-U(S)}$$

（前記領域S₂に含まれ且つ前記真偽をとる属性が真であるデータの数をV（S₂）、前記領域Sに含まれ且つ前記真偽をとる属性が真であるデータの数をV（S）、前記領域S₂に含まれるデータ数をU（S₂）、前記領域Sに含まれるデータ数をU（S）とする。）を第3の条件として前記領域切出手段に出力する手段とをさらに有する請求項10記載のデータ間結合ルール導出装置。

【請求項12】前記切り出された領域S内の各バケットのv(i，j)／u(i，j)が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性が真であるデータ

数の割合に等しくなるようv(i,j)を変更する手段と、当該変更されたv(i,j)及び入力された条件θiでもって、前記領域切出手段が動作するように命令する手段とを有する請求項10記載のデータ間結合ルール導出装置。

【請求項13】前記平面構成手段が、
複数の前記データから、X個のデータをランダムサンプリングする手段と、
サンプリングされたデータを各前記数値属性についてソートし、X・i／N（i＝1，2，・・N）番目に該当する数値及びX・n／M（n＝1，2，・・M）番目に該当する数値を記憶する手段と、
記憶された前記数値を基準にして、前記複数のデータをN×M個の前記バケットに入れる手段とを含む請求項10乃至12記載のデータ間結合ルール導出装置。

【請求項14】
【数11】

$$g(i,j)=v(i,j)-\theta u(i,j)$$

を評価値とし、
前記領域切出手段が、
前記平面の各列において、少なくとも1のバケットを含み且つ前記評価値が最大となるバケットの範囲を求める列内範囲導出手段と、
前記平面内の任意のバケット（座標(m,t)）の前列にあるもう1つのバケット（座標(m-1,l)）を含み且つ当該前列（第1列）までで最大の評価値を有する領域と、前記列内範囲導出手段により導出される、前記任意のバケット（座標(m,t)）及び前記もう1つのバケットと同一行であって前記任意のバケットと同一列にあるバケット（座標(m,l)）を含む評価値が最大となる範囲とを加えた領域を加算領域とする時、各前記任意のバケットに対し、それ自身及び当該任意のバケットを含む前記加算領域全域全体として評価値を最大化する前記もう1つのバケットを検出し、当該加算領域の評価値と共に記憶する検出記憶手段と、
前記平面全体で前記加算領域の評価値が最大となる前記任意のバケットを検出し、前記検出記憶手段により検出された前記もう1つのバケットを用いて、領域Sを導き出す手段とを含む請求項10記載のデータ間結合ルール導出装置。

【請求項15】切り出されるべき領域に含まれる最低限のデータ数である最小サポート数Uminを入力する手段と、
前記切り出された領域Sに含まれるデータ数U（S）と前記最小サポート数Uminと比較する手段と、
前記比較の結果、Umin≒U（S）であれば、当該領域Sを切り出されるべき領域として出力する手段と、
前記比較の結果、Umim＞U（S）又はUmim＜U（S）の場合には、新たな条件θiにて、前記領域切出手段が動作するように命ずる手段とを含む請求項10記載のデータ間結合ルール導出方法。

【請求項16】切り出されるべき領域における前記真偽をとる属性が真であるデータの数の割合minconfを入力する手段と、
前記切り出された領域Sが、minconf≒V（S）／U（S）（U（S）は前記領域Sに含まれるデータ数、V（S）は前記領域Sに含まれ且つ前記真偽をとる属性が真であるデータの数）である場合には、当該領域Sを出力する手段と、
mimconf＜V（S）／U（S）又はminconf＞V（S）／U（S）である場合には、新たな条件θiにて前記領域切出手段が動作するよう命ずる手段とを含む請求項10記載のデータ間結合ルール導出装置。

【請求項17】切り出された前記領域Sに対し、
【数12】

$$f(U(S),V(S))$$
$$=-V(S)\log\frac{V(S)}{U(S)}-(U(S)-V(S))\log\frac{U(S)-V(S)}{U(S)}$$
$$-(Vsum-V(S))\log\frac{Vsum-V(S)}{Usum-U(S)}$$
$$-(Usum-Vsum-U(S)+V(S))\log\frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}$$

（Usumは前記平面全体のデータ数、Vsumは前記平面全体に含まれる前記真偽をとる属性が真のデータの数）を計算し、その値を前記領域Sに対応して記憶するエントロピ計算手段と、
変更された条件θにて前記領域切出手段及び前記エントロピ計算手段が動作するように命ずる手段と、
前記エントロピ計算手段に記憶されたf（U（S），V（S））を最大化する領域Sを出力する手段とを含む請求項10記載のデータ間結合ルール導出装置。

【請求項18】切り出された前記領域Sに対し、
【数13】

$$f(U(S),V(S))=U(S)\left(\frac{V(S)}{U(S)}-\frac{Vsum}{Usum}\right)^2$$
$$+(Usum-U(S))\left(\frac{Vsum-V(S)}{Usum-U(S)}-\frac{Vsum}{Usum}\right)^2$$

（Usumは前記平面全体のデータ数、Vsumは前記平面全体に含まれる前記真偽をとる属性が真のデータの数）を計算し、その値を前記領域に対応して記憶するインタクラスバリアンス計算手段と、
変更された条件θにて前記領域切出手段と前記インタクラスバリアンス計算手段とを動作するように命ずる手段と、
前記インタクラスバリアンス計算手段に記憶されたｆ（Ｕ（Ｓ），Ｖ（Ｓ））を最大化する領域Ｓを出力する手段とを含む請求項１０記載のデータ間結合ルール導出装置。
【請求項１９】２種類の数値属性と、１種類の真偽をとる属性を含むデータを有するデータベースにおいて、コンピュータにデータ間の結合ルールを導き出させるプログラムコード手段を含む記憶装置であって、
コンピュータに、前記２種類の数値属性に対応する２つの軸を有し且つＮ×Ｍ個のバケットに分割されている平面の各バケットに対応して、当該バケット（座標(i, j)）に属するデータの数ｕ(i, j)及び前記真偽をとる属性が真であるデータの数ｖ(i, j)を記憶させる平面構成プログラムコード手段と、
コンピュータに、条件θを入力させる入力プログラムコード手段と、
コンピュータに、
【数１４】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大にするような前記バケットの領域Ｓを前記平面から切り出させる領域切出プログラムコード手段とを含む記憶装置。
【請求項２０】コンピュータ及び前記入力プログラムコード手段により、前記条件θとは異なる第２の条件θ_2を入力し、コンピュータ及び前記領域切出プログラムコード手段により、前記第２の条件θ_2に対応する第２の領域S_2を前記平面から切り出した場合に、
【数１５】

$$\theta 3=\frac{V(S_2)-V(S)}{U(S_2)-U(S)}$$

（前記領域S_2に含まれ且つ前記真偽をとる属性が真であるデータの数を$V(S_2)$、前記領域Ｓに含まれ且つ前記真偽をとる属性が真であるデータの数をＶ（Ｓ）、前記領域S_2に含まれるデータ数を$U(S_2)$、前記領域Ｓに含まれるデータ数をＵ（Ｓ）とする。）を第３の条件として前記領域切出プログラムコード手段に出力するプログラムコード手段とをさらに有する請求項１９記載の記憶装置。
【請求項２１】コンピュータに、前記切り出された領域Ｓ内の各バケットのｖ(i, j)／ｕ(i, j)が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性のデータ数の割合に等しくなるようｖ(i, j)を変更させるプログラムコード手段と、
当該変更されたｖ(i, j)及び入力された条件θ_4を用いて、コンピュータと前記領域切出プログラムコード手段が動作するように命じるコンピュータプログラムコード手段とを有する請求項１９記載の記憶装置。
【請求項２２】前記平面構成プログラムコード手段が、コンピュータに、複数の前記データから、Ｘ個のデータをランダムサンプリングさせるプログラムコード手段と、
コンピュータに、サンプリングされたデータを各前記数値属性についてソートし、Ｘ・ｉ／Ｎ（ｉ＝１，２，・・Ｎ）番目に該当する数値及びＸ・ｎ／Ｍ（ｎ＝１，２，・・Ｍ）番目に該当する数値を記憶させるプログラムコード手段と、
コンピュータに、記憶された前記数値を基準にして、前記複数のデータをＮ×Ｍ個の前記バケットに入れさせるプログラムコード手段とを含む請求項１９乃至２２記載の記憶装置。
【請求項２３】
【数１６】

$$g(i,j)=v(i,j)-\theta u(i,j)$$

を評価値とし、
前記領域切出プログラムコード手段が、
コンピュータに、前記平面の各列において、少なくとも１のバケットを含み且つ前記評価値が最大となるバケットの範囲を求めさせる列内範囲導出プログラムコード手段と、
前記平面内の任意のバケット（座標(m, t)）の前列にあるもう１つのバケット（座標(m-1, l)）を含み且つ当該前列（第ｌ列）までで最大の評価値を有する領域と、前記列内範囲導出プログラムコード手段とコンピュータにより導出される、前記任意のバケット（座標(m, t)）及び前記もう１つのバケットと同一行であって前記任意のバケットと同一列にあるバケット（座標(m, l)）を含む評価値が最大となる範囲とを加えた領域を加算領域とする時、コンピュータに、各前記任意のバケットに対し、

それ自身及び当該任意のバケットを含む前記加算領域全体として評価値を最大化する前記もう１つのバケットを検出させ、当該加算領域の評価値と共に記憶させる検出記憶プログラムコード手段と、
コンピュータに、前記平面全体で前記加算領域の評価値が最大となる前記任意のバケットを検出させ、前記検出記憶プログラムコード手段で検出された前記もう１つのバケットを用いて、領域Ｓを導き出させるプログラムコード手段とを含む請求項１９記載の記憶装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、データベースにおけるデータ相関の解析（データマイニングという。）に関し、より詳しくは２項の数値属性と１項の真偽をとる属性（真偽をとる条件又は０－１属性ともいう。）を有するデータ間の相関を見い出す手法に関する。

【０００２】

【従来の技術】例えば、銀行の顧客を解析対象とし、流動性預金残高がいくらくらいで且つ年齢が何歳ぐらいの人であれば、定期預金残高が２００万円以上になる人が全体の２０％となるか、といった問題を実際に解くことを考える。この流通性預金残高及び年齢は、整数ではあるが連続数値であり、一方定期預金残高２００万円以上というのは、２００万円以上か未満かという分類になるので、真偽をとる属性を有するものである。真偽をとる属性は、例えば「顧客がクレジットカードを有しているか」や「顧客が男性であるか」といった問題と置き換えることも可能である。このような課題を解決することができれば、銀行はどのような人に、例えば新型の金融商品に関するダイレクトメールを送ればよいか簡単に分かるので、効率的な営業活動が行える。

【０００３】従来、先に述べた真偽をとる属性間の相関を表現するルール（結合ルール、association rule）を高速に抽出するような研究は、データマイニングの分野において行われてきた。例えば、R.Agrawal, T.Imielinski, and A.Swami, "Miningassociation rules between sets of items in large databases" In proceedings of the ACM SIGMOD Conference on Management of data, May 1993. や、R.Agrawal and R.Srikant, "Fast algorithms for mining association rules" InProceedings of the 20th VLDB Conference, 1994. 等がある。

【０００４】また、２項の数値データ間のルールを求める従来手法には、以下のようなものがある。

１．強い線形相関を見い出すために、平面上の直線で、点集合を最適近似するものを探す方法。例えば、最小自乗法、再帰中央法等である。これら方法の欠点は、線形相関しか分からず、しかも相関係数の絶対値が０．５以下の場合に線形相関を用いて各データを予測すると精度が低く、現実にはほとんど役にたたない点にある。

２．弱い大域相関を見い出すためには、２次元平面上で正方形、長方形、又は円、楕円で面積に対して多くのデータを含むものを見い出す方法。例えば、計算幾何学アルゴリズムを利用するものである。この場合、計算時間が大きくなってしまうという欠点がある。例えば円の場合、Ｏ（M^3）以上の手間が掛かり得る（Ｏ（Ｍ）は、オーダーM^3の計算手間がかかることを示す。Ｍはデータ数である。）。また、取り出す相関領域としては決まった形をしたものしか扱うことができない。現実には、決まった形で適切にカバーできる場合は少ない。

３．平面を正方メッシュに分割しておき、たくさんのデータを含むピクセルを取り出す方法。しかし、取り出されたピクセルの集合は連結でなく、バラバラなことが多いので、ルールとして見い出すのは困難である。

【０００５】このような手法を用いると、上記の欠点の他に、データ間の多くのルールのうちで、意味のあるものと無意味なものとの区別が難しいという欠点もある。通常、相関に実用上の意味があるかどうかは人間の判断によらないといけないことが多いが、１．や２．では特殊な相関しか取り出せないので意味ある相関を見逃しやすく、３では出力を人間が見てルールを見い出せない。

【０００６】以上のように、従来２項以上の数値属性を有するデータ間の相関を見い出す手法に有効なものはなく、真偽をとる属性間のデータマイニングの手法を組み合わせることができるわけでもない。よって、最初に述べた例のような、２項以上の数値属性と真偽をとる属性を有するデータ間の相関を見い出すために用いることができる手法は現在のところ提案されていない。

【０００７】

【発明が解決しようとする課題】本発明は、以上のような点に鑑み、２項以上の数値属性と真偽をとる属性を有するデータ間の相関を見い出すための手法を提案するものである。

【０００８】特に、（１）真偽をとる属性が真であるデータの割合がある定められた値以上であって、含まれるデータ数が最大となるようなルールであるサポート最大化ルールや、（２）最低限含まれるデータ数が定めれた場合、真偽をとる属性が真であるデータの割合が最大となるようなルールであるコンフィデンス最大化ルール、（３）取り出される領域内部と外部との分割を考えた時に、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである最適化エントロピ・ルール、（４）領域内外の分割を考えた時に、内外の「標準化された真偽の割合の平均からのずれ」の二乗和を最大化するルールである最適化インタクラスバリアンス・ルールを満たすような範囲（領域）を導出可能とすることも目的である。

【０００９】さらに、上記のようなデータ間の相関を実時間内に行うことができるような手法を提供することも目的である。

【００１０】また、データ間の相関を人間に見やすい形

で提示することも目的である。そして、多くの相関の状態を可視化することにより、使用する人間の選択の幅を増大させることも目的とする。

【００１１】

【課題を解決するための手段】通常、解析対象物は多くの数値属性を有する。この中から２つの数値属性を選び、また、１つの真偽をとる属性について、以下のステップを行うことにより、上記の目的を達成するものである。すなわち、

（１）２つの数値属性により平面を構成し、この平面をメッシュ分割し、各メッシュ（バケットともいう）内のデータ数及び真偽をとる属性が真となったデータの数をカウントする。このような平面は、各メッシュをピクセルとした場合、データ数が濃淡度、真偽をとる属性が真となるデータの数が彩度に該当するような、平面画像として捉えることができる。

（２）所定の条件θに従い、平面の１つの軸に凸な領域である許容イメージを切り出し、データの相関の強い部分を見い出す。先に述べたように、平面を画像として捉え、平面の１つの軸に凸であるという条件を満たす部分画像（部分領域）を許容イメージとして切り出す。

（３）切り出した許容イメージとなる領域が、先に述べたようなサポート最大化ルール等の条件を満たしていれば、その領域をユーザに提示する。また、データベースからその領域に含まれるデータの必要な属性を引き出すことも、必要に応じて行う。

【００１２】なお、切り出された領域を、そのままユーザに提示したり、複数の領域を切り出した場合には、それを動画として可視化することにより、所望の結合ルールを見い出し易くすることもできる。

【００１３】また、一旦領域を切り出した後に、それ以外の結合ルールを見出すべく、切り出された領域について、彩度を平均化し、再度切り出しステップを実行することも可能である。

【００１４】最初に述べたような例の場合、流動性預金残高の軸と、年齢の軸を設け、その平面を適当なメッシュに分割する。そして、各メッシュについて該当する顧客の数と、定期預金残高２００万円以上の顧客の数をカウントする。そして、例えば顧客全体の２０％が入り且つ定期預金残高２００万円以上である顧客の割合が最大となるような許容イメージである領域の切り出しを行うことにより、コンフィデンス最大化ルールを得ることができる。

【００１５】また、例えば定期預金残高２００万円以上の顧客割合が１０％で最大の顧客数を有する領域を切り出すことにより、サポート最大化ルールを得ることができる。

【００１６】以上述べた事項をまとめると、２種類の数値属性と、１種類の真偽をとる属性を含むデータを有するデータベースにおいて、データ間の結合ルールを導き出す方法であって、２種類の数値属性に対応する２つの軸を有し且つＮ×Ｍ個のバケットに分割されている平面の各バケットに対応して、当該バケット（座標(i, j)）に属するデータの数ｕ(i, j)及び真偽をとる属性が真であるデータの数ｖ(i, j)を記憶する平面構成ステップと、条件θを入力するステップと、

【数１７】

$$\sum_{(i,j)\in S} g(i,j) = \sum_{(i,j)\in S} (v(i,j) - \theta u(i,j))$$

を最大にするようなバケットの領域Ｓを前記平面から切り出す領域切出ステップと、切り出された領域Ｓ内に含まれるデータを出力するステップとを含む。

【００１７】また、入力された前記条件θとは異なる第２の条件θ_2を入力するステップと、

【数１８】

$$\sum_{(i,j)\in S_2} g(i,j) = \sum_{(i,j)\in S_2} (v(i,j) - \theta_2 u(i,j))$$

を最大にするようなバケットの第２の領域S_2を平面から切り出すステップと、

【数１９】

$$\theta 3 = \frac{V(S_2) - V(S)}{U(S_2) - U(S)}$$

（領域S_2に含まれ且つ真偽をとる属性が真であるデータの数をＶ（S_2）、領域Ｓに含まれ且つ真偽をとる属性が真であるデータの数をＶ（Ｓ）、領域S_2に含まれるデータ数をＵ（S_2）、領域Ｓに含まれるデータ数をＵ（Ｓ）とする。）を第３の条件として、

【数２０】

$$\sum_{(i,j)\in S_3} = \sum_{(i,j)\in S_3} (v(i,j) - \theta_3 u(i,j))$$

を最大にするようなバケットの第３の領域S_3を平面から切り出すステップとを含むようにすることも考えられる。このような処理は、最初の条件θで、初期の目的のルールを導き出せなかった場合に有用である。

【００１８】さらに、切り出された領域Ｓ内の各バケットのｖ(i, j)、ｕ(i, j)が、平面全体のデータ数に対する平面全体の真偽をとる属性が真であるデータ数の割合に等しくなるようｖ(i, j)を変更するステップと、当該変更されたｖ(i, j)を用いて、入力された条件θ_4に従い、

【数２１】

$$\sum_{(i,j)\in S_4} g(i,j) = \sum_{(i,j)\in S_4} (v(i,j) - \theta_4 u(i,j))$$

を最大にするようなバケットの第４の領域S_4を切り出すステップとをさらに含むようにすることも考えられる。このようにすると、二次的な相関ルールを導き出すことができる。

【００１９】また、先の平面構成ステップが、複数のデータから、Ｘ個のデータをランダムサンプリングするステップと、サンプリングされたデータを各数値属性につ

いてソートし、X・i／N（i＝1，2，・・N）番目に該当する数値及びX・n／M（n＝1，2，・・M）番目に該当する数値を記憶するステップと、記憶された数値を基準にして、複数のデータをN×M個のバケットに入れるステップとを含むようにすることも考えられる。このようにすると、各バケットにデータを高速に割り振ることができる。

【００２０】ここで、

【数２２】

$$g(i,j)=v(i,j)-\theta u(i,j)$$

を評価値とし、先の領域切出ステップが、平面の各列において、少なくとも１のバケットを含み且つ評価値が最大となるバケットの範囲を求める列内範囲導出ステップと、平面内の任意のバケット（座標(m,t)）の前列にあるもう１つのバケット（座標(m-1,l)）を含み且つ当該前列（第１列）までで最大の評価値を有する領域と、列内範囲導出ステップにより導出される、前記任意のバケット（座標(m,t)）及び前記もう１つのバケットと同一行であって前記任意のバケットと同一列にあるバケット（座標(m,l)）を含む評価値が最大となる範囲とを加えた領域を加算領域とする時、各前記任意のバケットに対し、それ自身及び当該任意のバケットを含む加算領域全体として評価値を最大化する前記もう１つのバケットを検出し、当該加算領域の評価値と共に記憶する検出記憶ステップと、平面全体で加算領域の評価値が最大となる前記任意のバケットを検出し、検出記憶ステップで検出＊された前記もう１つのバケットを用いて、領域Ｓを導き出すステップとを含むようにすることができる。

【００２１】また、切り出されるべき領域に含まれる最低限のデータ数である最小サポート数Uminを入力するステップと、切り出された領域Ｓに含まれるデータ数Ｕ（Ｓ）と最小サポート数Uminと比較するステップと、この比較の結果、Umin≒Ｕ（Ｓ）であれば、当該領域Ｓを切り出されるべき領域として出力するステップと、前記比較の結果、Umim＞Ｕ（Ｓ）又はUmim＜Ｕ（Ｓ）の場合には、新たな条件θ_sにて、先の領域切出ステップを実施するステップとを含むようにすることが考えれる。これにより、コンフィデンス最大化ルールを導出することができる。

【００２２】さらに、切り出されるべき領域における真偽をとる属性が真であるデータの数の割合minconfを入力するステップと、切り出された領域Ｓが、minconf≒Ｖ（Ｓ）／Ｕ（Ｓ）（Ｕ（Ｓ）は領域Ｓに含まれるデータ数、Ｖ（Ｓ）は領域Ｓに含まれ且つ真偽をとる属性が真であるデータの数）である場合には、当該領域Ｓを出力するステップと、mimconf＜Ｖ（Ｓ）／Ｕ（Ｓ）又はminconf＞Ｖ（Ｓ）／Ｕ（Ｓ）である場合には、新たな条件θ_sにて領域切出ステップを実施するステップとを含むようにすることも考えられる。これによりサポート最大化ルールを導出することができる。

【００２３】ここで、切り出された前記領域Ｓに対し、

【数２３】

$$f(U(S),V(S)) = -V(S)\log\frac{V(S)}{U(S)} - (U(S)-V(S))\log\frac{U(S)-V(S)}{U(S)} - (Vsum-V(S))\log\frac{Vsum-V(S)}{Usum-U(S)} - (Usum-Vsum-U(S)+V(S))\log\frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}$$

（Usumは平面全体のデータ数、Vsumは平面全体に含まれる真偽をとる属性が真のデータの数）を計算し、その値を領域Ｓに対応して記憶するエントロピ計算ステップと、条件を変更して先の領域切出ステップとエントロピ計算ステップを実行するステップと、ｆ（Ｕ（Ｓ），Ｖ※（Ｓ））を最大化する領域Ｓを出力するステップとを含むようにすれば、最適化エントロピ領域を見つけることができる。

【００２４】また、切り出された前記領域Ｓに対し、

【数２４】

$$f(U(S),V(S)) = U(S)\left(\frac{V(S)}{U(S)}-\frac{Vsum}{Usum}\right)^2 + (Usum-U(S))\left(\frac{Vsum-V(S)}{Usum-U(S)}-\frac{Vsum}{Usum}\right)^2$$

を計算し、その値を領域に対応して記憶するインタクラスバリアンス計算ステップと、条件を変更して先の領域切出ステップとインタクラスバリアンス計算ステップとを実行するステップと、ｆ（Ｕ（Ｓ），Ｖ（Ｓ））を最大化する領域Ｓを出力するステップとを含むようにすることも考えられる。これにより、最適化インタクラスバリアンス領域を切り出すことができる。

【００２５】以下の説明を理解すれは、上述の方法を実施するような装置を作成すること、またこのような方法をコンピュータに実施させるプログラムを格納した記憶

15

装置又は記憶媒体を作成することは容易に実施できるであろう。

【００２６】

【発明の実施の形態】まず、本発明の各ステップがどのように実施されるかを示す。

（１）平面構成ステップ

図１に、平面構成ステップのフローを示す。ステップ１００にて処理が開始し、まずデータ集合Ｐからデータのランダムサンプリングを行う（ステップ１１０）。そして、サンプリングされたデータｐｉ（ｘｉ，ｙｉ）（ｘｉ，ｙｉはデータの２つの数値属性の値を示す。）のｘｉ，ｙｉごとにソートを行う（ステップ１２０）。この時、２つの数値属性の各々に対応するように２軸を有する平面を考え、その平面を各軸ごとにＮ個のバケットに分割する。すなわち、平面上にはN^2個のバケットが存在することとなる。そして、ｘｉ，ｙｉごとに、ｉ・Ｘ／Ｎ（Ｉ＝１，２，．．．Ｎ－１）番目の値を見つけ出す（ステップ１３０）。Ｘはサンプリングされたデータの数である。このようにすると、平面上の各列及び各行に属するデータの数は一致する。そして、ｘｉ，ｙｉごとのｉ・Ｘ／Ｎ（Ｉ＝１，２，．．．Ｎ－１）番目の値を用いると、各バケットの境界数値が分かるので、それを用いて各バケット（座標(i,j)）に入るデータｐｉの数ｕ(i,j)と、真偽をとる属性が真であるデータの数ｖ(i,j)を、各バケットごとにカウントする（ステップ１４０）。各バケットごとに、ｕ(i,j)とｖ(i,j)を記憶する（ステップ１５０）。言い換えると、ｕ(i,j)行列とｖ(i,j)行列が生成されたということができる。

【００２７】上述のようにランダムサンプリングを行うのは、全てのデータをソートしていると時間がかかるからである。但し、ソートしてもよい場合もある。また、ランダムサンプリングで取り出されるデータの数は、３０Ｎから５０Ｎぐらいが好ましい。また、２軸ともＮ個に分割する例を示したが、異なる数に分割することも可能である。典型的な例で、Ｎは１００ぐらいである。

【００２８】以上述べたのは一例であって、他の方法を用いてもよい。例えば、各バケットの境界数値については予め定めた値を用いても良い。また、データ値に対して均等に分割することも、また対数的に分割することも可能である。

【００２９】また、後の処理のため以下のような処理（図２）を行っておくと、さらに全体の処理が高速化される。すなわち、ｕ(i,j)とｖ(i,j)の行数（N_y）と列数（N_x）を調べる（ステップ２１０）。そして、先に求めたｕ(i,j)とｖ(i,j)を用いて、新たに以下のようなｕ'(i,j)とｖ'(i,j)という行列を作成する（ステップ２２０）。

【数２５】

$$u'(i,j)=\sum_{k=0}^{i} u(k,j)$$

16

【数２６】

$$v'(i,j)=\sum_{k=0}^{i} v(k,j)$$

【００３０】このｕ'(i,j)とｖ'(i,j)は、後々数多く計算することとなる目的関数、

【数２７】

$$g(i,j)=v(i,j)-\theta u(i,j)$$

の和計算を以下のように簡単化するために用意する。

【数２８】

$$\sum_{k=i}^{j} g(k,m)=\sum_{k=i}^{j} v(k,m)-\theta\sum_{k=i}^{j} u(k,m)$$
$$=(v'(j,m)-v'(i-1,m))-\theta(u'(j,m)-u'(i-1,m))$$

【００３１】さらに、

【数２９】

$$Usum=\sum_{全体} u$$

【数３０】

$$Vsum=\sum_{全体} v$$

も後によく用いるので用意する。

【００３２】以上のような準備をすれば、以下の領域切り出しステップが高速になる。なお、上述のように作成された平面は、平面"画像"と考えることもできるので、イメージという文言を用いる場合がある。同様に、バケットはピクセルということもある。

【００３３】（２）領域切り出しステップ

このステップにおいては許容イメージである領域を先に作成した平面（平面画像）から切り出すものである。この許容イメージは、先に述べたように１つの軸方向に凸であるイメージを言う。より正確に言うと、Ｘ軸方向へ単調な２本の曲線で囲まれる連結イメージを言う。このことは、図３に例を示す。左には、幅１のＹ軸方向に伸びる帯で切ると必ず連結している例を示しており、このようなイメージを許容イメージという。また、右には、先の帯で切ると連結していない例が示されており、このようなイメージを含めて切り出そうとすると、その問題はＮＰ困難となってしまう。

【００３４】このような許容イメージに問題を限定すると、Ｙ軸方向に伸びる帯は必ず連結しているので、ダイナミックプログラミングを用いて、それらを順々に連結していけばよい。但し、もう１つのパラメータを指定する必要がある。例えば、濃淡のみの画像の場合にはピクセル数（バケット数）でもよい。例えばユーザにピクセル数を指定させ、そのピクセル数で且つ濃淡度の総和が最大になる許容イメージを取り出すことができる。

【００３５】全体の許容イメージの連結性は、以下の事項を考えると分かる。すなわち、ｋ個のピクセル（バケ

ット）を含み、m列より左のピクセルからなり、（m，t）の位置を含むイメージ（図4参照のこと）を考え、この時に濃淡値の総和をf（k，m，t）とすると、こ＊れは以下の数式を満たすものである。

【数31】

$$f(k,m,t) = \max_{t,l \in I}\{f(k-|I|,m-1,l)+\sum_{i \in I} q(m,i)\}$$

【0036】ここで、IはX軸上m列目の連続区間（範囲）であり、t，lを含む。q(i,j)は、座標(i,j)の濃淡度を示す。また、f（k－|I|，m－1，l）を計算する時に用いたX軸上m－1列目の連続区間（範囲）はlを含むので、Iと連結している。従って、帰納法よりf（k，m，t）を求めるまでに用いた連続区間全体は連結していることが保証されている。

【0037】このようにすると、O（N^6）のオーダーで計算が可能となり、$N^2=n$（nはピクセル数）とするとO（n^3）となり、実時間で計算可能となる。

【0038】以上濃淡のみの画像においてピクセル数（バケット数）を特定する場合の例を説明したが、本発明では濃淡のみではないので、他の方法を用いる。すなわち、図5に示すような、横軸が切り出される領域Sに含まれるデータ数U（S）、縦軸が切り出される領域Sに含まれ且つ真偽をとる属性が真であるデータの数V（S）であるような平面を考える。データ数と真偽をとる属性が真であるデータの数の組み合わせは多数存在するので、この平面には多数の点が存在することになるが、この点のうち、凸包を構成する点を特に用いる。すなわち、この凸包を構成する点をつなぐことにより曲線を構成し、この曲線に対し傾きθを有する直線を上から下ろして行き、最初にこの曲線と接する点を求め、この時のイメージを出力するという方法を用いる。この時の許容イメージをフォーカス・イメージ（focused image）という。フォーカス・イメージを図5では黒丸で示している。また、直線を下ろしていくような方法をハンドプローブという。このように、本発明では傾きθを入力するような方法を用いる。

【0039】このように凸包上の点のみ取り扱うのは、※コンフィデンス最大化ルール、サポート最大化ルールは、凸包上に必ず存在するわけではないが、近似解としては十分な点を出力することができ、また最適化エントロピ・ルール及び最適化インタクラスバリアンス・ルールについては、この凸包上に必ず存在するからである。もし、コンフィデンス最大化ルール及びサポート最大化ルールの厳密解を解くとすると、実時間には計算が終了しないので、近似解であっても十分に有効な結果を計算できる。

$$f(m,t) = \max_{t,l \in I}\{f(m-1,l)+\sum_{i \in I} g(i,m)\}$$

$$= \max_{l}\{f(m-1,l)+\underbrace{\max_{t,l \in I}\overbrace{\sum_{i \in I} g(i,m)}^{B}}_{A}\}$$

この数33のAは、t，lを含む連続区間（範囲）全体で数33のBを最大化する連続区間（範囲）Iを見つけることを意味する。

【0042】このIをcover（t，l）と記述することとする。いま、t≦lを仮定すると、次の定義されるlow（t），high（l）を用いれば、

【数34】

$$cover(t,l) = [low(t),t] \cup [t,l] \cup [l,high(l)]$$

となる。但し、low（t）は連続区間［i，t］全体で数33のBが最大となるiであり、high（l）は連続区間［l，j］全体で数33のBが最大となるjを言う。

【0043】このlow（t）やhigh（l）はダイナミック・プログラミング中では何度も用いられるの

【0040】上記のように傾きθの直線を下ろしていくということは、直線y＝θx＋QのY切片であるQを減少させることであり、言いかえれば、Q＝V（S）－θU（S）を最大にするU（S）をX座標に有する点を求める問題となる。よって、

【数32】

$$\max Q = \max\{\sum_{(i,j) \in S} v(i,j) - \theta \sum_{(i,j) \in S} u(i,j)\}$$

$$= \max \sum_{(i,j) \in S} g(i,j)$$

と変形される。

【0041】では、この数32をどのように解くかを考える。基本的には先に述べたダイナミック・プログラミングを用いた手法を用いる。まず、m列目より左のバケットからなり、（m，t）の位置のバケットを含む領域を考え、この中で目的関数である数27を最大化したものをf（m，t）とすると、以下の条件を満たす　すなわち、

【数33】

19

で、ｌｏｗ（ｔ）やｈｉｇｈ（ｌ）を高速に求めることができれば有効である。このため、連続区間［ｉ，ｊ］中の数３３のＢが要素Ｋ（ｉ，ｊ）に入る行列Ｋを作る。但し、ｉ＞ｊの場合にはＫ（ｉ，ｊ）＝（ｉ－ｊ）ｘ（ｘは十分、その絶対値が大きい負の数。例えば、ｕ（ｉ,ｊ)を全体について加算した値より絶対値が大きいならば十分である。）としておく。すると行番号ｌにおいて最大値を有する列の列番号がｈｉｇｈ（ｌ）となる。よって、すべてのｌ∈［１，Ｎ］についてｈｉｇｈ（ｌ）を求める問題は、Ｋの各行の最大値の列番号を求める問題となる。このような計算は、Ｏ（Ｎ）の計算量で行える。

【００４４】最大値の列番号を求める場合に最大値かどうかを判断するための行列の要素の比較を行う。この比較は、先に数２５及び数２６にて求めたｕ’及びｖ’を用いると簡単に済む。

【００４５】このＫで各行の最大値に注目すると、行番号が大きくなるにつれて最大値の列番号は単調に増加する。但し、同じ行に最大値が複数個ある時には、左端のみを考える。このような行列を「単調な行列」と呼ぶ。＊

20

＊証明は省略するがこの行列Ｋは「完全単調な行列」（任意の部分行列が単調行列であるような行列。）でもある。Ｋの一例を図６に示し、斜線部が各行の最大値である。単調な行列の全ての行の最大値を有する列番号を計算するにはＯ（ＮｌｏｇＮ）の計算量が必要である。この完全単調な行列の各行の最大値の列番号を求めるアルゴリズムは周知であり、例えば、「計算幾何学」浅野哲夫著，朝倉書店，１９９０年９月の第４章「計算幾何学の基本的技法」に記載されている。

【００４６】同様にしてｌｏｗ（ｔ）を計算する場合には、連続区間［ｉ，ｊ］中の数３３のＢが要素Ｌ（ｊ，ｉ）に入る行列Ｌを作り、行番号ｔにおける最大値を有する列の列番号を求めれば、ｌｏｗ（ｔ）になる。今度は上三角部分（ｉ＞ｊ）を－∞とする。このような行列も完全単調な行列である。

【００４７】このようにして求めたｌｏｗ（ｔ）及びｈｉｇｈ（ｌ）を用いれば、ｃｏｖｅｒ（ｔ，ｌ）が求まり、数３３の変形である以下の式が計算可能となる。

【数３５】

$$f(m,t)=\max_{\ell}\{f(m-1,\ell)+\sum_{i\in cover(t,\ell)}g(i,m)\}$$

【００４８】ｆ（ｍ，ｔ）を最大にするような、イメージを見い出すには、ｆ（ｍ，ｔ）の計算をＸ軸に垂直な帯について順に行い、その帯を記憶しておき、それらを連結すれば求まる。

【００４９】さらに高速化するには、

【数３６】

$$M(t,\ell)=f(m-1,\ell)+\sum_{i\in cover(t,\ell)}g(i,m)$$

を要素として有する行列Ｍを作り、行番号ｔの最大値がｆ（ｍ，ｔ）となる。Ｍは先に説明した完全単調な行列であり、すべてのｔについてｆ（ｍ，ｔ）はＯ（Ｎ）で計算することができる。よって、すべてのｍについてｆ（ｍ，ｔ）を計算するには、Ｏ（Ｎ^2）の計算量が必要となる。

【００５０】以上詳細を述べたが、必要なステップを以下に示しておく。

（１）全てのＸ軸に垂直な帯についてｌｏｗ（ｔ），ｈｉｇｈ（ｌ）を計算しておく。

（２）ｌｏｗ（ｔ），ｈｉｇｈ（ｌ）によりｃｏｖｅｒ（ｔ，ｌ）が求まるので、数３６を要素とする行列Ｍを計算する。

（３）行列Ｍの各行の最大値を求め、その値をｆ（ｍ，ｔ）として記憶する。

（４）イメージ全体を把握するために、行列Ｍの各行の列番号ｌをｓ（ｍ，ｔ）に入力する。

（５）ｆ（ｍ，ｔ）を最大にするｍ，ｔを求め、（４）で作ったｓ（ｍ，ｔ）及びｓ（ｍ，ｔ）に記憶されているｌを用いて前列のｌｏｗ（ｔ），ｈｉｇｈ（ｌ）でもって、イメージを把握する。

【００５１】このステップを図７及び図８に示す。ステップ６１０で開始した処理は、ステップ６２０においてｍ＝１の初期化を行う。そしてｍ＝Ｎｘ－１であるかを判断する（ステップ６３０）。これはループをＮｘ回繰り返すためであり、Ｎｘはｕ(ｉ,ｊ)とｖ(ｉ,ｊ)の列数である。

【００５２】この後に、

【数３７】

$$K(i,j)=\begin{cases}(i,j)x & (i>j)\\ \sum_{h\in[i,j]}g(h,m) & (i\leqq j)\end{cases}$$

を計算しておき、この行列の各行の最大値を求め、その列番号ｊをｈｉｇｈ（ｍ，ｉ）とする（ステップ６４０）。これにより、前記平面（平面画像）のｍ列目のｈｉｇｈ（ｉ）が求まった。

【００５３】また、

【数３８】

$$L(i,j)=\begin{cases}-\infty & (j>i)\\ \sum_{h\in[j,i]}g(h,m) & (j\leqq i)\end{cases}$$

を計算しておき、この行列の各行の最大値を求め、その列番号ｊをｌｏｗ（ｍ，ｉ）＝ｊとする（ステップ６５０）。これにより、前記平面（平面画像）のｍ列目のｌｏｗ（ｉ）が求まった。

【００５４】その後にｍを１インクメントして（ステッ

プ６６０）、ステップ６３０に戻る。このように、まずｌｏｗとｈｉｇｈを最初にすべて計算しておく。図８の計算を実行するごとに必要なｌｏｗとｈｉｇｈを計算するようにしてもよいが、この例のように一度に計算してしまってもよい。上述のように一度に計算した後に処理はＸを介して図８に進む。

【００５５】図８においてＸから、ｆ（ｍ，ｔ）を計算する。まず、図４のような前記平面の一番左の列について処理する。すなわち、

【数３９】

$$f(1,t) = \sum_{i \in [low(1,t),\ high(1,t)]} g(i,1)$$

をｔ＝１からＮｙについて計算する（ステップ７１０）。Ｎｙはｕ(i,j)とｖ(i,j)の行数である。このようにすると、数３６の第１項の初期値となるものが計算されたこととなる。

【００５６】また、ｓ（１，ｔ）＝－１としておく、これは、最初の列であるから、これ以上前の列には連結があり得ないことを明示するためである。

【００５７】そして、ｍ＝２以降の値を計算するために、ｍ＝２とし（ステップ７２０）、以下のループをＮｘ－１回まわすため、ｍ＝Ｎｘ＋１かどうか判断する（ステップ７３０）。もし、ｍ＝Ｎｘ＋１でなければ、ｆ（ｍ－１，ｉ）（$1 \leqq i \leqq N_y$）に負の値が一つでもある場合には、ｔ＝１からN_yまで、

【数４０】

$$f'(m,t) = \sum_{i \in cover_m(t,t)} g(i,m)$$

を計算する（ステップ７４２）。ここで、ｃｏｖｅｒｍ（ｔ，ｔ）は、ｍ列目のｃｏｖｅｒ（ｔ，ｔ）の意味である。そして、

【数４１】

$$M(t,\ell) = f(m-1,\ell) + \sum_{i \in cover_m(t,\ell)} g(i,m)$$

この行列Ｍの各行の最大値の列番号を求める（ステップ７４４）。但し、その列ｉの最大値とｆ’（ｍ，ｉ）と比較して大きい方をｆ（ｍ，ｉ）とする。そして、ｓ（ｍ，ｔ）には、ｆ’（ｍ，ｔ）がＭのｔ行の最大値よりも小さいときには１を、そうでなければ－１を入れる。これは、前列までの連結を保存した方が切り出される領域の目的関数の値が大きくなるか、それとも前列までの連結を放棄した方が目的関数の値が大きくなるかという判断をしているのである。

【００５８】このようにｓ（ｍ，ｔ）は連結状態を保存するためにあるので、ある（ｍ，ｔ）が決まれば、ｓ（ｍ，ｔ）を手繰って遡り、領域がどのように連結するものかを後にみることができる。

【００５９】この後に、ｍを１インクリメントし（ステップ７５０）、ステップ７３０に戻る。繰り返しが全て終われば、ｆ（ｍ，ｔ）を最大とするｍ，ｔが求まる。ｆ（ｍ，ｔ）を作りながら、常に最大となるｍ，ｔを保持しておき、新たに作成された部分につき保持しているｍ，ｔより大きな点を見い出した場合には更新するようにしていけばよい。このｍ，ｔを用いて、ｓ（ｍ，ｔ）から、前列のｌが求まる。このｌとｔのうち小さい方をｌｏｗに大きい方をｈｉｇｈに入力する。例えば、ｔの方が小さければ、ｌｏｗ（ｍ，ｔ），ｈｉｇｈ（ｍ，ｌ）が求まる。また、ｓ（ｍ－１，ｌ）から、さらに前列のｌ’が求まるので、ｌｏｗ（ｍ－１，ｌ），ｈｉｇｈ（ｍ－１，ｌ’）を求める。（ここでは、ｌの方がｌ’より小さい。）これを繰り返していくと、イメージＳの全体が分かる（ステップ７６０）。あるｓ（ｍ，ｔ）にて－１が得られれば、その領域は終了する。これにて、傾きθを入力した場合に、そのθに対応する許容イメージＳを得ることができたので、処理を終了する（ステップ７７０）。

【００６０】以上説明したように、傾きθに対応する許容イメージを得ることができた。この条件θは、ユーザが入力するようにしてもよいし、また例えばサポート最大化ルールを求めるためにシステムにより設定された条件であってもよい。

【００６１】（３）出力ステップ

以上のように求まった許容イメージたる領域Ｓは、前記平面のどの部分を占めているかは、先のステップによりわかっているので、そのイメージＳに属するデータを取り出すことになる。通常各データは、真偽をとる属性及び数値属性のみならず、他の属性も有しているから、例えばダイレクトメールを送るのであれば、住所氏名といった属性を取り出すようになる。ここまでくると、取り出すべきデータは特定されているから、通常のデータベースの検索に過ぎないので、これ以上詳しく述べない。当然、一旦フォーカス・イメージをその外形がよくわかるようにして、ユーザに提示するようにしてもよい。

【００６２】以上のような各ステップを実施すれば、ある条件θに対する、データ間結合ルールの１つを求めることができる。しかし、この条件θをどのように設定するかということは、１つの問題である。通常、ある条件θ１つでは、問題の解決にならない場合が多い。以上の各ステップ、特に（１）平面構成ステップと（２）領域切り出しステップとをエンジンとして用い、どのように先に述べた４つの一般的なルール及び他のルール等を導き出すかを以下に示す。

【００６３】Ａ．ある区間に存在するフォーカス・イメージを求める場合

まず、幾つかのθに対応する許容イメージ（ここではフォーカス・イメージ）たる領域Ｓを連続的に示し、動画を作成することにより、切り出されるイメージの大きさ

及び形状をユーザの判断により決定させる場合を考える。

【0064】この処理を図9に示す。ステップ800にて開始された処理は、まずθ1を入力することにより、上述したプロセスにてフォーカス・イメージS1を見つけ出す（ステップ810）。また、ユーザにθ2を入力させ、同様にフォーカス・イメージS2を見つけ出す（ステップ820）。このようにして2つのフォーカス・イメージが求まると、それぞれに含まれるデータ数U（S1），U（S2）及び真偽をとる属性が真であるデータの数V（S1），V（S2）とを用いて、その中間にある、新たな傾きθ3を計算する（ステップ830）。

【0065】このように新たなθ3が求まれば、さらにこのθ3に対応するフォーカス・イメージS3を求めることができる（ステップ840）。ここで、計算されたS3が既に求まっていれば、区間（θ1，θ2）にはこれ以上のフォーカス・イメージは存在しない。よって、処理が終了する（ステップ880）。しかし、発見済みでなければ、θ2の代わりにθ3を用いて、ステップ830以降を実行する（ステップ860）。すなわち、区間（θ1，θ3）の間にあるフォーカス・イメージを見つけ出す。この場合、次々に中間の値を計算していくようにすることも可能である。また、ある程度の個数フォーカス・イメージが求まったところで計算を取り止めることもできる。さらに、もう1つ残った区間（θ3，θ2）についてフォーカス・イメージを計算するために、θ3，θ2についてステップ830以降を実行する（ステップ870）。この場合も、この区間内に存在しているフォーカス・イメージを全て見つけ出すようにしてもよいし、所定の個数見つけ出したところで処理を終了してもよい。

【0066】このようにして、1つ又は複数のフォーカス・イメージを見つけ出すことができた。このように求まった複数のフォーカス・イメージを図10のように（a），（b），（c）と連続してユーザに提示するようなことも可能である。この図10の各々の斜線部分が切り出された領域（フォーカス・イメージ）である。

【0067】B．コンフィデンス最大化ルールの場合（図11及び図12）

この場合には、ルールの定義より最小限度のサポートminsup（全体のデータ数に対する包含されるデータ数の割合）を入力することとなる（ステップ910）。ここで、Umin＝Usum*minsupを計算しておく。ここで図5をみてみると、最小限度サポートと記された縦の点線がこの値に対応する。そして、θ＝Vsum／Usumとしてフォーカス・イメージS1を求める（ステップ920）。このS1に含まれるデータ数U（S1）の値により3つの場合に分けられる（ステップ930）。すなわち、

（1）U（S1）≒Uminであれば、当然S1が解として出力され、処理は終了する（ステップ990）。（2）U（S1）＜Uminであれば、θ＝0としてフォーカス・イメージS2を計算する（ステップ940）。この結果、U（S2)≒Uminであれば（ステップ950）、当然S2が解として出力され、処理は終了する。ここで、S1＝S2はminsupが1より小さければあり得ない。但し、S1＝S2であれば解はないので、解なしを出力する。そうでなければ、XXにて図12に移動する。

（3）U（S1）＞Uminであれば、θ＝1としてフォーカス・イメージS2を計算する（ステップ960）。S1＝S2であるならば、S1より高いコンフィデンスを有するフォーカス・ポイントは存在しないので、S1（当然S2も同様）を最良解として出力し、処理を終了する（ステップ970）。同様に、U（S2）≒Uminであれば、S2を出力して処理を終了する。ステップ970にて、先に示した条件を満たさないとされた場合には、XXを介して図12に移動する。

【0068】図12では、XXから始まり、新たな条件θを求め、このθに対するフォーカス・イメージSを計算する（ステップ1400）。このθは

θ＝（V（S2）－V（S1））／（U（S2）－U（S1））

にて計算される。そして、S1＝S又はS2＝Sであるならば、（S1，S2）の間にはこれ以上フォーカス・イメージは存在しないので、コンフィデンスの高いS2が最良解として出力され、処理を終了する（ステップ1410）。また、U（S）≒Uminであるならば、Sを出力し、処理を終了する。

【0069】ところが、U（S）＜Uminであると（ステップ1420）、まだ処理が必要なので、S1＝Sとして（ステップ1430）、ステップ1400に戻る。同様に、U（S）＞Uminであるならば、S2＝Sとして（ステップ1440）、ステップ1400に戻る。

【0070】これを繰り返すことにより解が見つけられる。図5を参照すると、先に説明した最小限度のサポートの右側、濃く塗られた部分に解の存在する範囲がある。そして、この図5の場合には、凸包の内部の白丸の点が厳密解となるが、本発明ではハンド・プローブにて得られた近似解が出力される。見つけられた解は、ユーザに提示されるようにしてもよいし、そのフォーカス・イメージに属するデータの必要な属性を出力するようにしてもよい。

【0071】C．サポート最大化ルールの場合（図13，図14）

このルールの場合、定義より、最小限度のコンフィデンス、すなわち含まれるデータ数に対する真偽をとる属性が真である割合を入力する（ステップ1110）。図5の場合、最小限度のコンフィデンスと示され、原点から引かれた点線がこれに該当する。次に、minconf≦Vsum／Usumであるかを判断する（ステップ1120）。この条件に合致する場合には、全ての凸包上の点は条件に

合致するので、θ＝0としてフォーカス・イメージを求め、それを出力する。（ステップ１１３０）。これに対し、ステップ１１２０でminconf＞Vsum／Usumである場合には、θ＝Vsum／Usumとしてフォーカス・イメージS1を求める（ステップ１１４０）。ここで、3つの場合が考えられる（ステップ１１５０）。すなわち、（1）minconf≒V（S）／U（S1）である場合には、S1を出力して処理を終了する（ステップ１１９０）。（2）minconf＜V（S1）／U（S1）である場合には、θ＝0のフォーカス・イメージS2を求める（ステップ１１６０）。このS2は、V（S2）＝Vsumであり、且つU（S2）＝Usumであるから、minconf≦V（S2）／U（S2）とはなり得ない。同様に、S1＝S2はあり得ない。よって、図１４への移動するためにYに移る（ステップ１１６０）。（3）minconf＞V（S1）／U（S1）である場合には、θ＝minconfのフォーカス・イメージS2を求める（ステップ１１７０）。このS2に対し、minconf＝V（S2）／U（S2）が成り立てば、S2は厳密解であるから、これを出力して処理を終了する。また、minconf＜V（S2）／U（S2）であれば、解は存在しないので、解なしを返して処理を終了する。一方、minconf＞V（S2）／U（S2）であるならば、図１４の処理に移行するためにYに進む。

【００７２】図１４では、Yから処理が開始され、θ＝（V（S2）－V（S1））／（U（S2）－U（S1））としてフォーカス・イメージSを求める（ステップ１２００）。この求められたフォーカス・イメージSに対し、（1）minconf≒V（S）／U（S）が成立する場＊合には、このSを出力して処理を終了する（ステップ１２１０）。また、S1＝S若しくはS2＝Sである場合には、これ以上S1とS2の間には解は無いので、S1を最良解として出力し、処理を終了する（ステップ１２１０）。これに対し、minconf＜V（S）／U（S）である場合には（ステップ１２２０）、S1＝Sとしてステップ１２００に戻る（ステップ１２３０）。また、minconf＞V（S）／U（S）である場合には、S2＝Sとしてステップ１２００に戻る（ステップ１２４０）。

【００７３】以上のようにして、サポート最大化ルールが求められる。もう一度図５に戻ると、先に説明した最小限度のコンフィデンスとして示した点線より上の濃く塗られた範囲に解が存在する。そして、この例では凸包内の白丸の点が厳密解であるが、このように凸包内部の点は見つけ出すのに膨大な計算量を必要とするので、凸包上の点でサポートを最大にする近似解を出力するようにしている。先に述べたように、見出された近似解又は厳密解は、ユーザに提示してもよいし、フォーカス・イメージ内に含まれるデータの必要な属性値を出力するようにしてもよい。

【００７４】D．最適化エントロピ・ルールの場合

最適化エントロピ・ルールとは、領域の内部と外部との分割を考えた時、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである。よって、切り出された領域と平面全体のエントロピのゲイン（以下の式）が最大となる領域を発見すればよい。

【数４２】

$$f(x,y) = -y\log\frac{y}{x} - (x-y)\log\frac{x-y}{x} - (b-y)\log\frac{b-y}{a-x}$$
$$-(a-b-x+y)\log\frac{a-b-x+y}{a-x}$$

このxはU（S）、yはV（S）、aはUsum、bはVsumである。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、θを変化させ、数４２を最大化するフォーカス・イメージを求めればよい。

【００７５】E．最適化インタクラスバリアンス・ルールの場合

※先に述べたように最適化インタクラスバリアンス・ルールとは、領域内外の分割を考えた時、内外の「標準化された真偽の割合の平均からのずれ」の二乗和を最大化するルールである。よって、切り出された領域と平面全体のインタクラスバリアンス（以下の式）が最大となる領域を発見すればよい。

【数４３】

$$f(x,y)=x\left(\frac{y}{x}-\frac{b}{a}\right)^2+(a+x)\left(\frac{b-a}{a-x}-\frac{b}{a}\right)^2$$

x，y，a，bは上述したものと同じである。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、θを変化させ、数４３を最大化するフォーカス・イメージを求めればよい。

【００７６】F．その他

以上述べたように、U（S）とV（S）上の凸包上の点に存在する又は存在すると近似できる場合には、上述したステップを用いれば高速にルールに該当する領域を導き出すことができる。

【００７７】G．二次的なルールの抽出

上述のプロセスを用いて１つのルールを見い出した後に、二次的なルールを見つけ出すことができる。すなわち、切り出した１のフォーカス・イメージに属するv

(i,j)を除去し、v(i,j)′u(i,j)＝Vsum′Usumとなるように、v(i,j)を変更し、それから新たに領域切り出しステップを行うのである。

【００７８】以上、本発明における処理のプロセスを説明した。このような処理プロセスは、コンピュータ・プログラムによって実現し、実行するようにしてもよい。例えば、図１５のような通常のコンピュータ・システムにおいて実行できるようなプログラムにすることもできる。処理プログラムは、ＨＤＤ１０５０に格納され、実行時にはメインメモリ１０２０にロードされ、ＣＰＵ１０１０によって処理される。また、ＨＤＤ１０５０はデータベースをも含んでおり、処理プログラムはそのデータベースに対するアクセスを行う。最初の平面やフォーカス・イメージ（図１７）は、表示装置１０６０によってユーザに提示される。ユーザは、入力装置１０７０にてフォーカス・イメージの選択や、データ出力の命令を入力する。このような入力装置には、キーボードやマウス、ポインティング・デバイスやディジタイザを含む。さらに、出力結果を補助記憶装置であるＦＤＤ１０３０のフロッピー・ディスクに記憶したり、また新たなデータをＦＤＤ１０３０から入力することもできる。さらに、ＣＤ－ＲＯＭドライブ１０４０を用いて、データを入力することもできる。

【００７９】さらに、本発明の処理プロセスを実現したコンピュータ・プログラムは、フロッピー・ディスクやＣＤ－ＲＯＭといった記憶媒体に記憶して、持ち運ぶことができる。この場合、通常のデータベース検索プログラムのデータ取り出し部分や、表示装置１０６０に表示するだけの処理を行うプログラムは、すでにＨＤＤ１０５０に記憶されている場合もある。よって、それ以外の部分が、上記のような記憶媒体にて流通することは通常行われる事項である。

【００８０】また、本発明の処理を専用に行うような装置を設けてもよい。例えば、図１６のような装置が考えられる。平面構成部１３１０は、データベース１３００及び切出部１３２０に接続されており、制御部１３４０からの命令を受付ける。また、切出部１３２０は、出力部１３３０及び表示装置１３５０に接続されており、制御部１３４０からの命令を受付ける。また、切出部１１３０はデータベースにも接続を有している。制御部１３４０は、入力部１３６０に接続され、入力部１３６０により指示された処理の種類により平面構成部１３１０及び切出部１３２０を制御する。

【００８１】この装置の簡単な動作を説明する。平面構成部１３１０は、先に説明した平面構成ステップを実行する部分である。このように平面構成部１１１０は、データベースに記憶されたデータを用いて先に示した平面を構成し、切出部１３２０に出力する。切出部１３２０は、制御部１３４０からの命令に従って、切り出しのためのパラメータであるθをセットする。セットされたθに従って切出部１３２０は、先に述べた切出ステップを行い、フォーカス・イメージを切り出す。そして、表示装置１１４０に出力し、ユーザに命令されれば、切り出されたフォーカス・イメージ内に属するデータをデータベース１３００から取り出し、出力部１３３０に引き渡す。出力部１３３０は、適当な形式でユーザ所望のデータを出力する。また、ユーザは、例えば入力部１３６０からコンフィデンス最大化ルールを解くように命じ、最小限度のサポートを入力する。すると、制御部１３４０は先に示した処理Ｂを行うように、条件θを設定し、切出部１３２０に出力する。そして、命じられたコンフィデンス最大化ルールに合致するような領域を解くべく、条件θを変化させる等の処理を行う。先に述べたサポート最大化ルール（処理Ｃ）や、最適化エントロピ・ルール（処理Ｄ）、最適化インタクラスバリアンス・ルール（処理Ｅ）、その他凸包上に位置する領域を切出す処理Ｆに適した条件θを切出部１３２０に渡す処理を制御部１３４０は行う。ユーザは入力部１３６０から処理の種類や、先に述べたような条件（θのみならず、mincon f, minsupも）を入力する。また、制御部１３４０は、上述の処理Ｇを行うために平面構成部１３１０に、切り出したフォーカス・イメージのV(i,j)を除去する等の処理を命じる。

【００８２】以上、本発明を専用の装置にする一例を示したが、本発明はこれに限定されるものではない。例えば、切出部１３２０の出力は、出力制御部を介して出力部１１６０及び表示装置１１４０に出力されるようにしてもよいし、この場合出力制御部からデータベースを参照してデータを取り出すようにしてもよい。

【００８３】以上は、通常データが有するＫ個の数値属性のうち２項を選択し、それらの数値属性間の壮観を見つける処理であったが、数２７を目的関数とし、ｎ次元空間の領域を切り出すことができれば、ｎ次元の探索に拡張することができる。

【００８４】

【効果】以上述べたように、２項の数値属性と真偽をとる属性を有するデータ間の相関を見い出すことができた。

【００８５】（１）サポート最大化ルール、（２）コンフィデンス最大化ルール、（３）最適化エントロピ・ルール、（４）最適化インタクラスバリアンス・ルールを満たすような範囲（領域）を導出可能とすることもできた。

【００８６】さらに、上記のようなデータ間の相関を実時間内に行うこともできた。

【００８７】また、データ間の相関を人間に見やすい形で提示することもできた。

【００８８】例えば、ある割合以上で、例えばアウトドアスポーツに興味を示す（真偽をとる属性に相当する）、できるだけまとまった領域に入る顧客を知ること

29

ができるので、その条件に合致する多くの顧客に知ってもらいたいダイレクトメールの宛て先を知るのに用いることができる。（サポート最大化ルール）

【００８９】一定数以上の顧客を含む、例えば定期預金残高２００万円以上の顧客割合が最も高いところを知ることができるので、顧客を絞りこみつつ、有効な宣伝活動等を行うことができる。（コンフィデンス最大化ルール）

【図面の簡単な説明】

【図１】平面構成ステップのフローを示す図である。

【図２】領域切り出しステップのための前準備のフローを示す図である。

【図３】許容イメージを説明するための図である。

【図４】許容イメージの連結性を説めいするための図である。

【図５】Ｕ（Ｓ），Ｖ（Ｓ）平面の説明をするための図である。

【図６】完全単調な行列の説明をするための図である。

【図７】切出ステップの一部を示すフローチャートである。

【図８】切出ステップの一部を示すフローチャートである。

【図９】複数のフォーカス・イメージを見つけ出す処理のフローを示す図である。

30

＊【図１０】フォーカス・イメージの一例を示した図である。

【図１１】コンフィデンス最大化ルールを導出するための処理の一部を示すための図である。

【図１２】コンフィデンス最大化ルールの導出するための処理の一部を示すための図である。

【図１３】サポート最大化ルールの導出するための処理の一部を示すための図である。

【図１４】サポート最大化ルールの導出するための処理の一部を示すための図である。

【図１５】通常のコンピュータ・システムで本発明を実施した場合の装置構成の一例を示す図である。

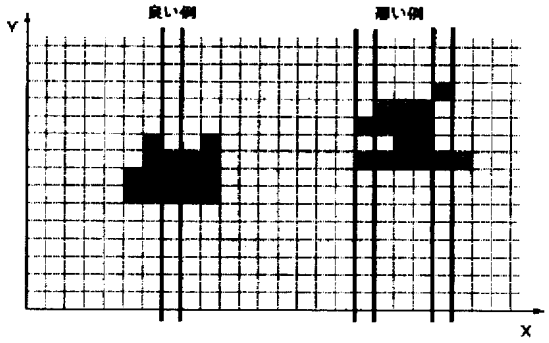
【図１６】本発明を専用の装置で実施した場合のブロック図である。

【図１７】図１５の表示装置の表示例を示す図である。

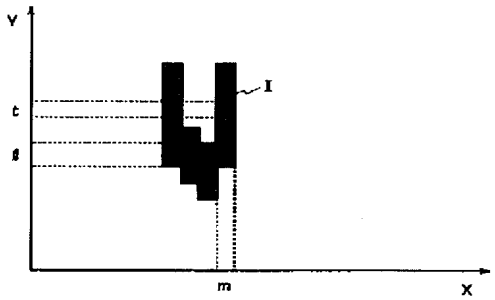
【符号の説明】

１０１０	ＣＰＵ	１０２０	メインメモリ
１０３０	ＦＤＤ	１０４０	ＣＤ－ＲＯＭドライブ
１０５０	ＨＤＤ	１０６０	表示装置
１０７０	入力装置		
１３１０	濃淡画像構成部		
１３００	データベース	１３２０	切出部
１３５０	表示装置	１１３０	入力部
１３３０	出力部	１３４０	制御部

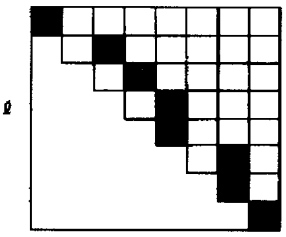
【図３】



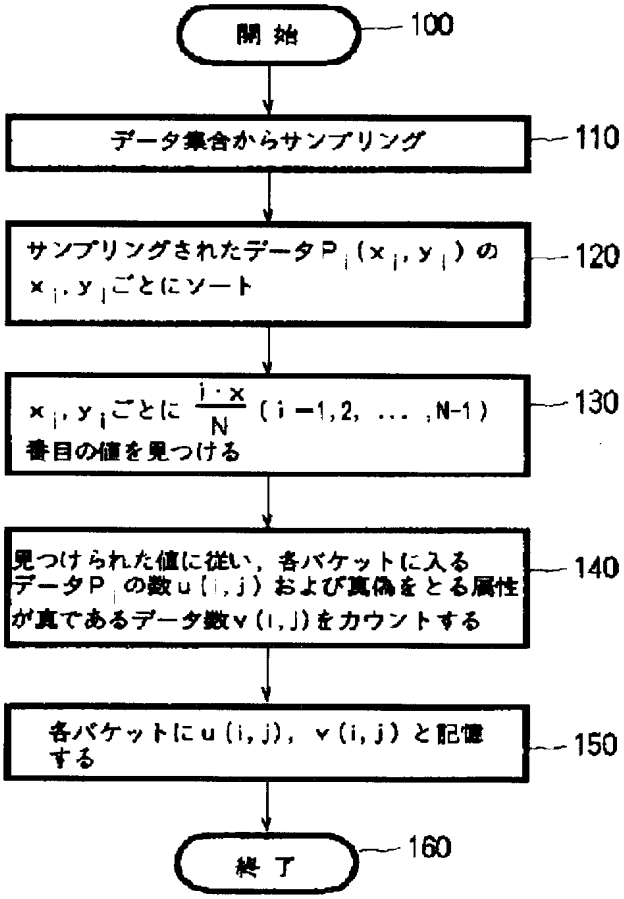
【図４】



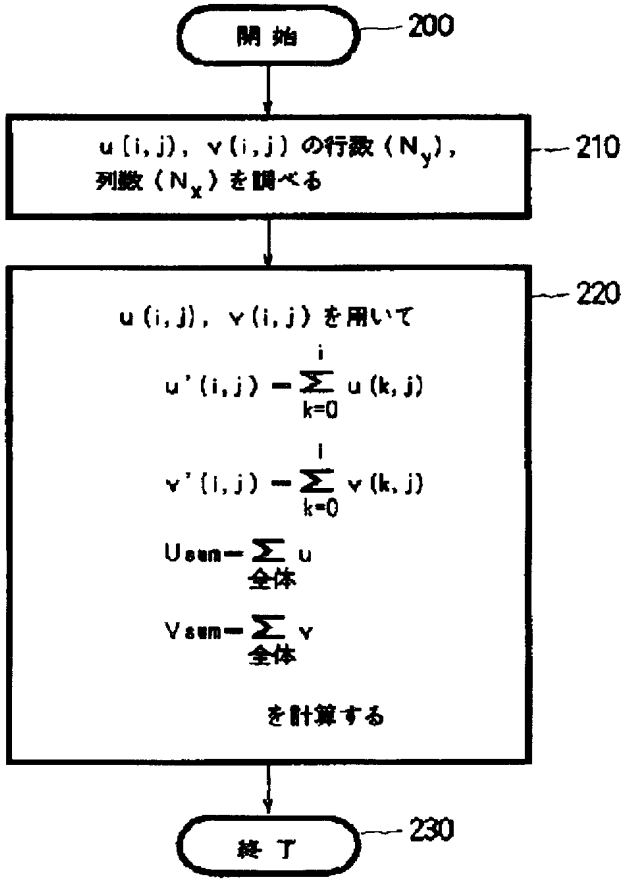
【図６】



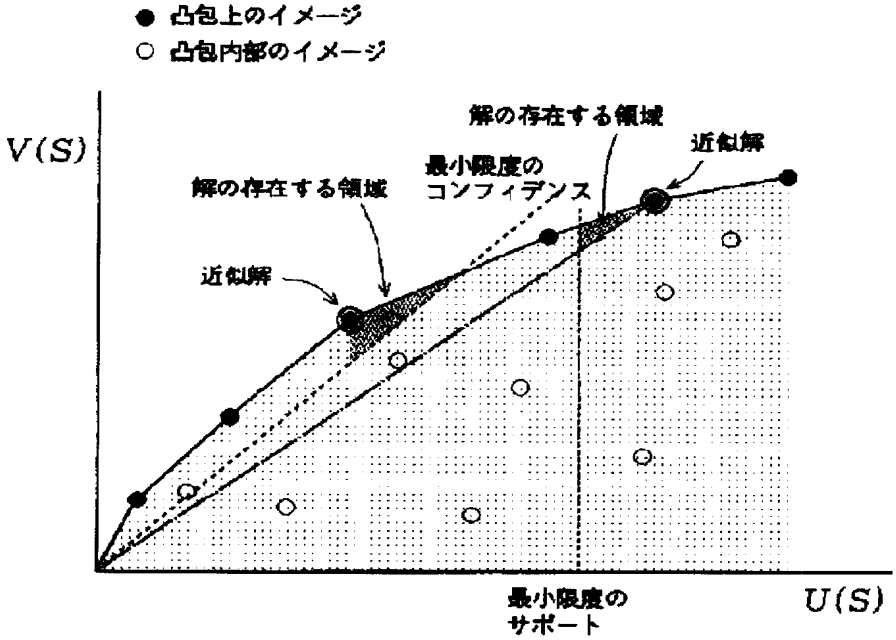
【図1】


開始
100
データ集合からサンプリング
110
サンプリングされたデータ$P_i(x_i, y_i)$のx_i, y_iごとにソート
120
x_i, y_iごとに $\frac{j \cdot x}{N}$ (i=1, 2, ..., N-1) 番目の値を見つける
130
見つけられた値に従い，各バケットに入るデータP_iの数u(i, j)および真偽をとる属性が真であるデータ数v(i, j)をカウントする
140
各バケットにu(i, j)，v(i, j)と記憶する
150
終了
160

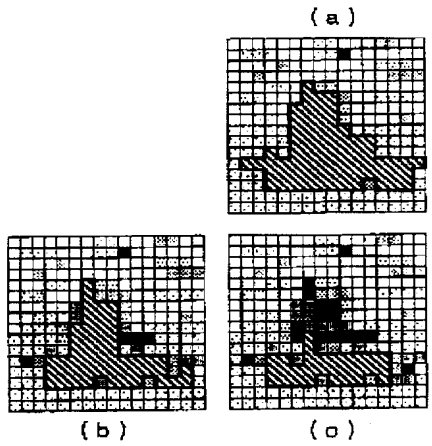
【図2】


開始
200
u(i, j)，v(i, j)の行数(N_y)，列数(N_x)を調べる
210
u(i, j)，v(i, j)を用いて
$u'(i,j) = \sum_{k=0}^{i} u(k,j)$
$v'(i,j) = \sum_{k=0}^{i} v(k,j)$
$Usum = \sum_{全体} u$
$Vsum = \sum_{全体} v$
を計算する
220
終了
230

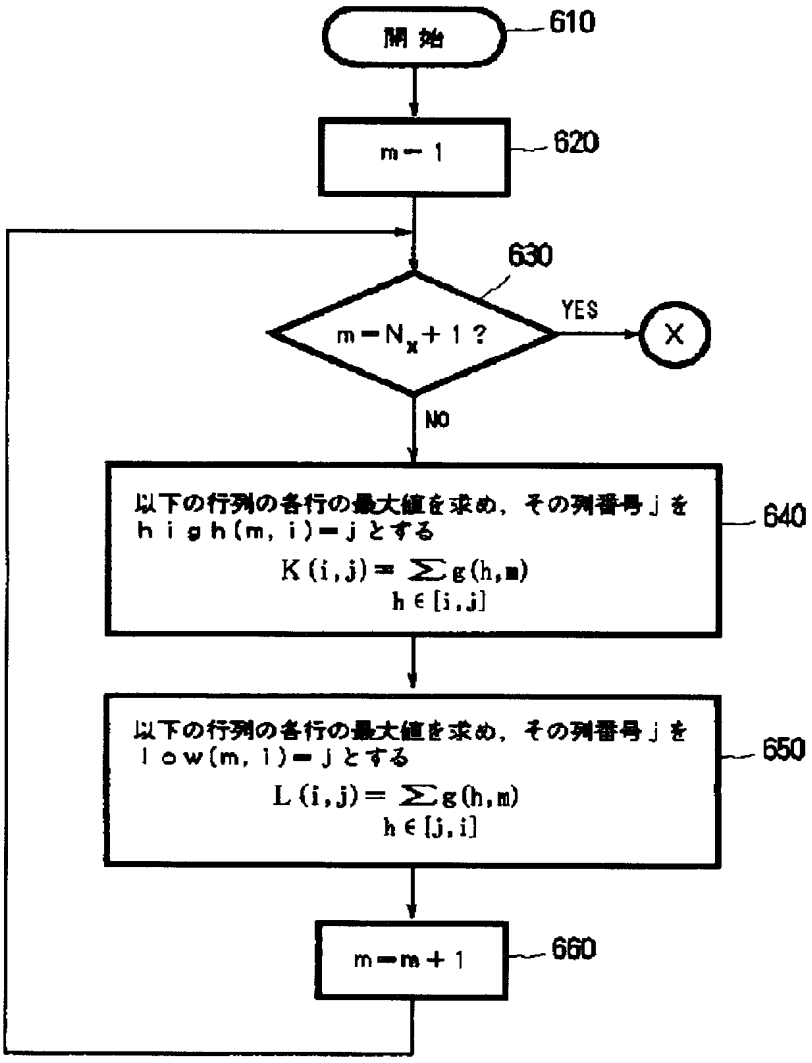
【図5】


● 凸包上のイメージ
○ 凸包内部のイメージ
V(S)
解の存在する領域
最小限度の
コンフィデンス
近似解
解の存在する領域
近似解
最小限度の
サポート
U(S)

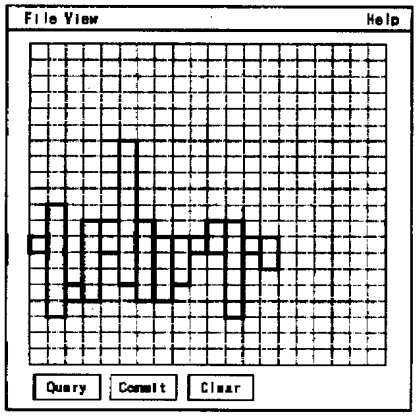
【図10】


(a)
(b)
(c)

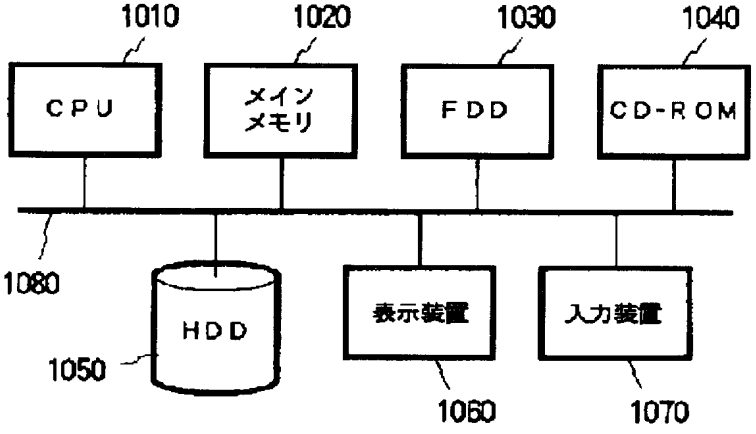
【図７】


開始
610
m－1
620
630
m－N_x＋1？
YES
X
NO
以下の行列の各行の最大値を求め，その列番号ｊを
ｈｉｇｈ(m，ｉ)－ｊとする
K(i,j)＝Σg(h,m)
h∈[i,j]
640
以下の行列の各行の最大値を求め，その列番号ｊを
ｌｏｗ(m，ｉ)－ｊとする
L(i,j)＝Σg(h,m)
h∈[j,i]
650
m－m＋1
660

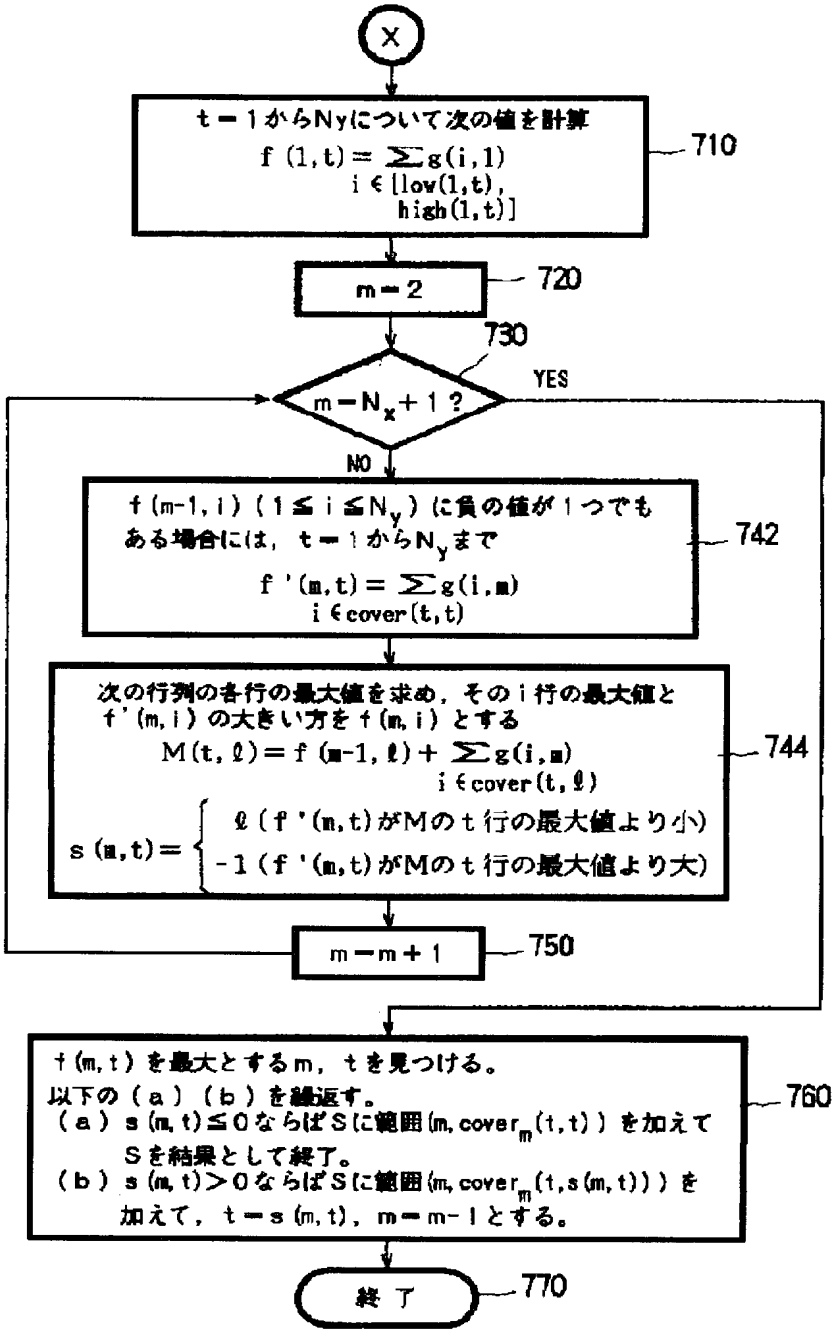
【図１７】


File View
Help
Query
Commit
Clear

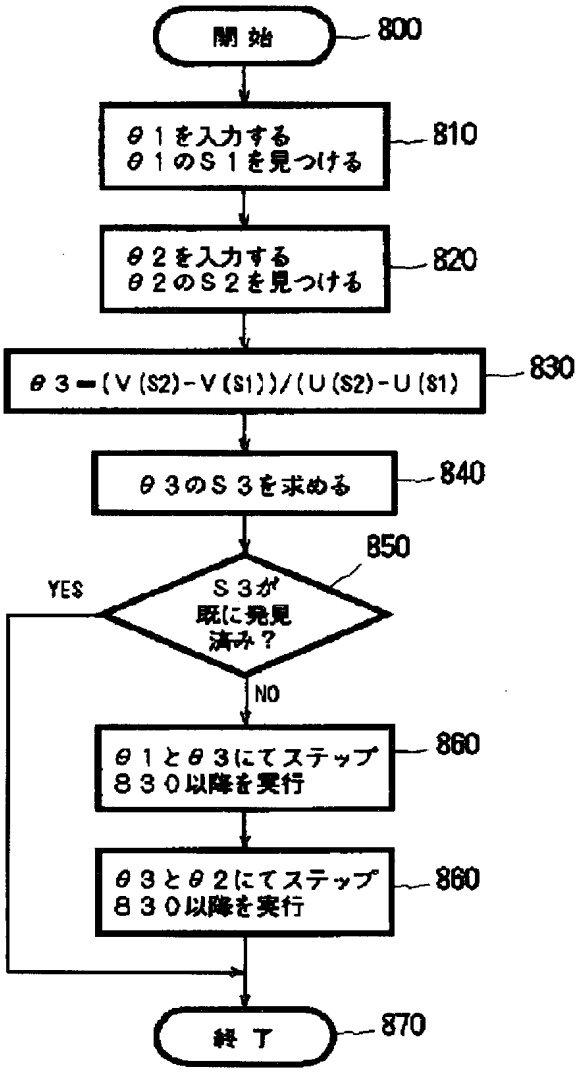
【図１５】


1010
1020
1030
1040
CPU
メイン
メモリ
FDD
CD-ROM
1080
HDD
1050
表示装置
1060
入力装置
1070

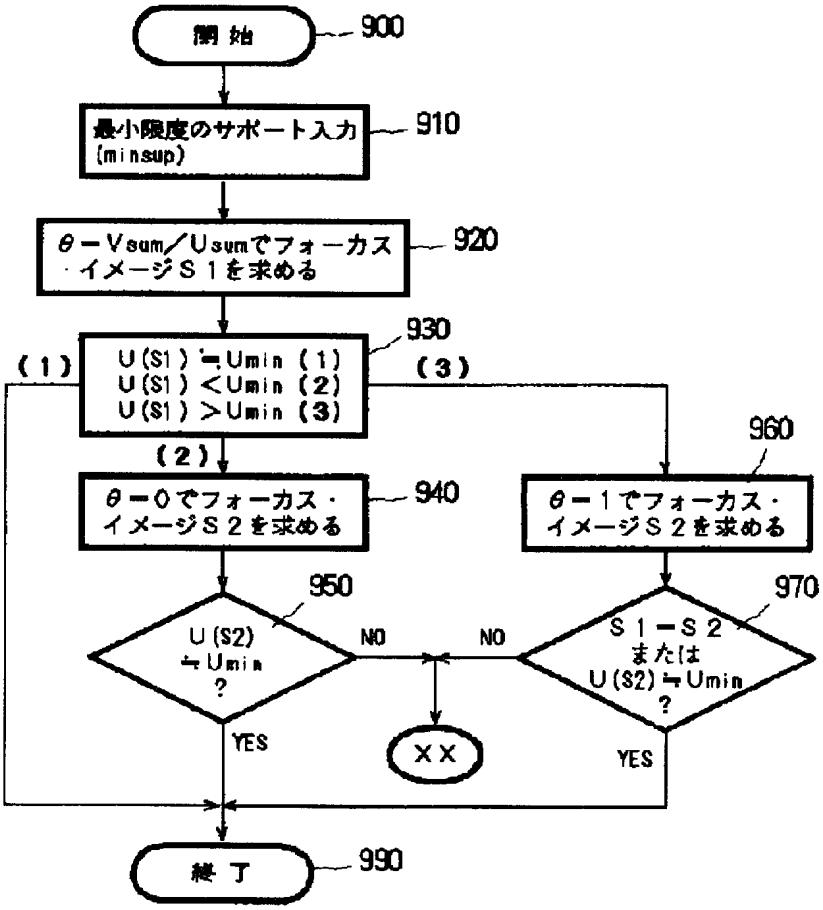
【図８】


X
t＝１からNyについて次の値を計算
f(1,t)＝Σg(i,1)
i∈[low(1,t), high(1,t)]
710
m＝2
720
730
m＝Nx＋１？
YES
NO
f(m-1,i)（１≦i≦Ny）に負の値が１つでもある場合には，t＝１からNyまで
f'(m,t)＝Σg(i,m)
i∈cover(t,t)
742
次の行列の各行の最大値を求め，そのi行の最大値とf'(m,i)の大きい方をf(m,i)とする
M(t,ℓ)＝f(m-1,ℓ)＋Σg(i,m)
i∈cover(t,ℓ)
s(m,t)＝ ℓ（f'(m,t)がMのt行の最大値より小）
-1（f'(m,t)がMのt行の最大値より大）
744
m＝m＋１
750
f(m,t)を最大とするm，tを見つける。
以下の（a）（b）を繰返す。
（a）s(m,t)≦０ならばSに範囲(m,cover_m(t,t))を加えてSを結果として終了。
（b）s(m,t)＞０ならばSに範囲(m,cover_m(t,s(m,t)))を加えて，t＝s(m,t)，m＝m-１とする。
760
終了
770

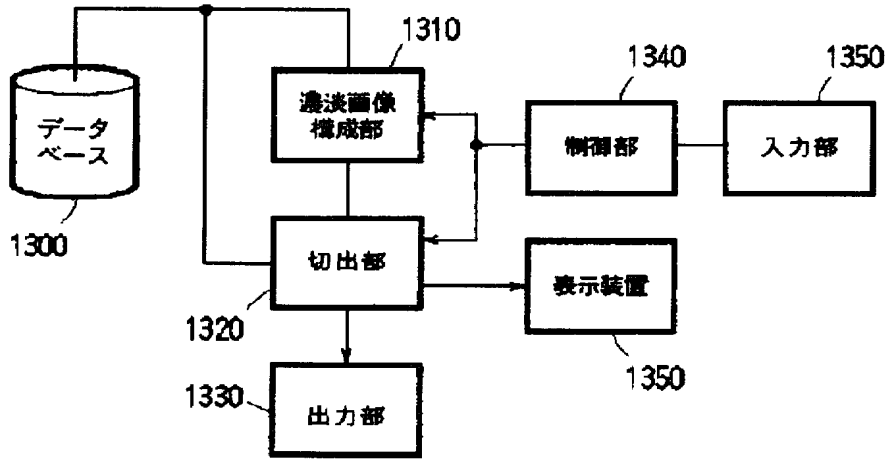
【図9】


開始
800
θ1を入力する
θ1のS1を見つける
810
θ2を入力する
θ2のS2を見つける
820
θ3＝(V(S2)-V(S1))/(U(S2)-U(S1)
830
θ3のS3を求める
840
S3が
既に発見
済み？
850
YES
NO
θ1とθ3にてステップ
830以降を実行
860
θ3とθ2にてステップ
830以降を実行
860
終了
870

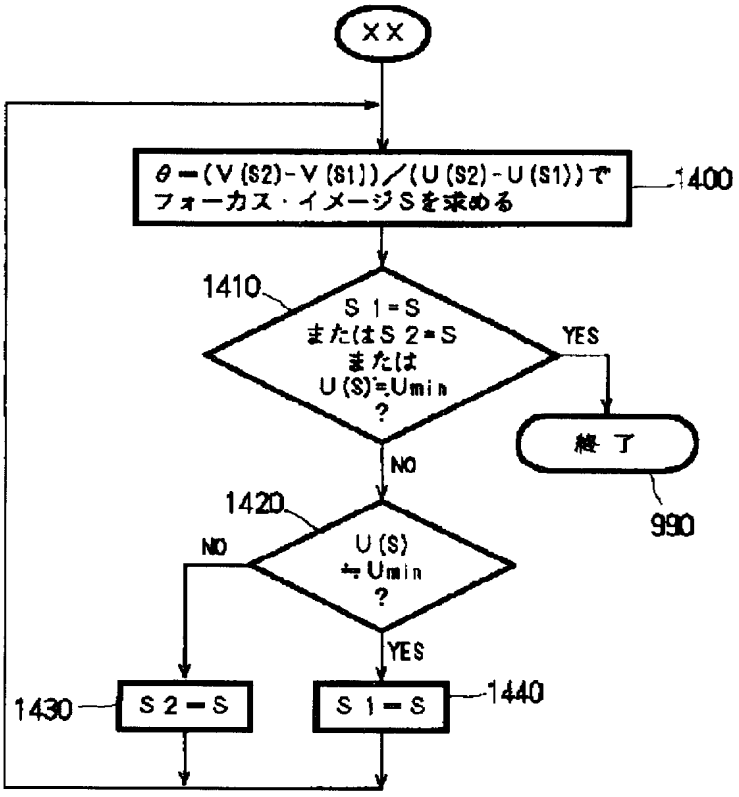
【図１１】


開始
900
最小限度のサポート入力
(minsup)
910
θ－Vsum／Usumでフォーカス
・イメージＳ１を求める
920
930
U(S1) ≒ Umin (1)
U(S1) ＜ Umin (2)
U(S1) ＞ Umin (3)
(1)
(2)
(3)
θ－0でフォーカス・
イメージＳ２を求める
940
θ－1でフォーカス・
イメージＳ２を求める
960
950
U(S2)
≒ Umin
?
NO
NO
S１－Ｓ２
または
U(S2) ≒ Umin
?
970
YES
YES
XX
終了
990

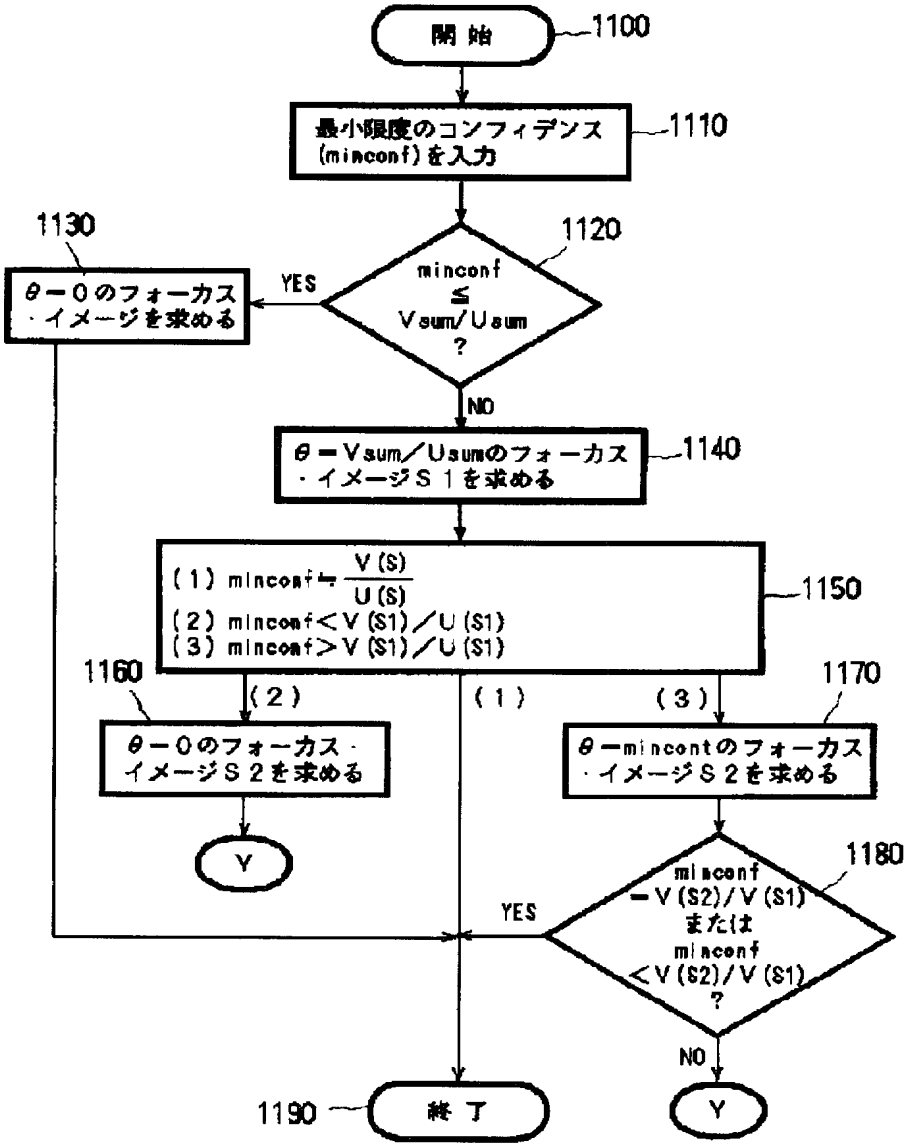
【図１６】


データ
ベース
1300
濃淡画像
構成部
1310
制御部
1340
入力部
1350
切出部
1320
表示装置
1350
出力部
1330

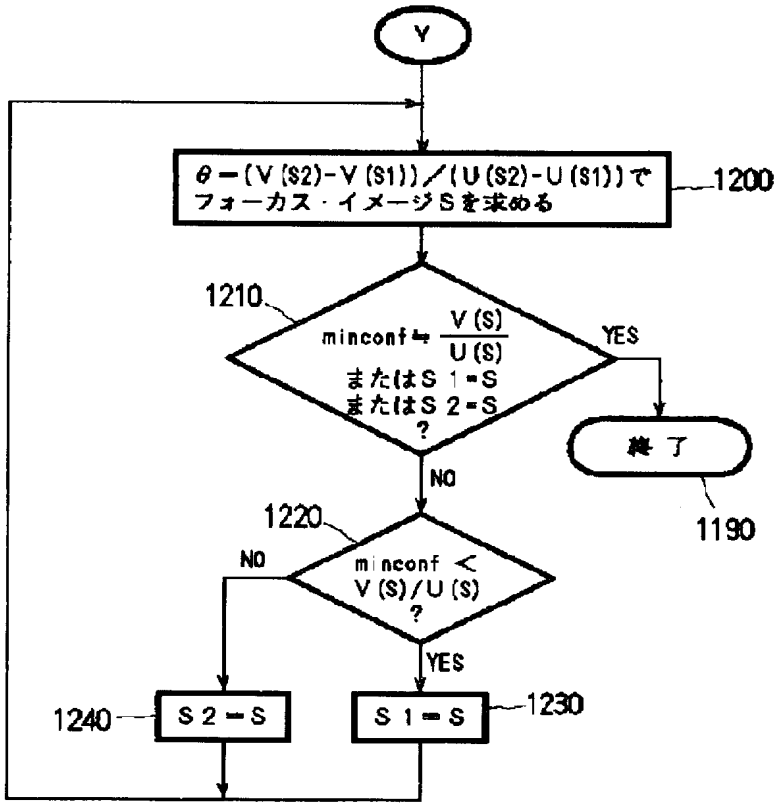
【図１２】


XX
θ＝(V(S2)-V(S1))／(U(S2)-U(S1))で
フォーカス・イメージSを求める
1400
1410
S1=S
またはS2=S
または
U(S)≒Umin
?
YES
終了
990
NO
1420
U(S)
≒Umin
?
NO
YES
1430
S2=S
S1=S
1440

【図１３】


開始
1100
最小限度のコンフィデンス (minconf) を入力
1110
minconf ≦ Vsum/Usum ?
1120
YES
θ＝0のフォーカス・イメージを求める
1130
NO
θ＝Vsum／Usumのフォーカス・イメージＳ１を求める
1140
(1) minconf≒ V(S)/U(S)
(2) minconf＜V(S1)／U(S1)
(3) minconf＞V(S1)／U(S1)
1150
(2)
(1)
(3)
1160
θ＝0のフォーカス・イメージＳ２を求める
1170
θ＝minconfのフォーカス・イメージＳ２を求める
Y
minconf ＝V(S2)/V(S1) または minconf ＜V(S2)/V(S1) ?
1180
YES
NO
Y
終了
1190

【図１４】



【手続補正書】
【提出日】平成８年６月５日
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】請求項８
【補正方法】変更
＊【補正内容】
【請求項８】切り出された前記領域Ｓに対し、
【数７】

＊

$$f(U(S),V(S)) = -\frac{V(S)}{Usum}\log\frac{V(S)}{Usum} - \frac{U(S)-V(S)}{Usum}\log\frac{U(S)-V(S)}{Usum} - \frac{Vsum-V(S)}{Usum-U(S)}\log\frac{Vsum-V(S)}{Usum-U(S)} - \frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}\log\frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}$$

（Ｕsumは前記平面全体のデータ数、Ｖsumは前記平面全体に含まれる前記真偽をとる属性が真のデータの数）を計算し、その値を前記領域Ｓに対応して記憶するエントロピ計算ステップと、
条件θを変更して前記領域切出ステップと前記エントロピ計算ステップを実行するステップと、
ｆ（Ｕ（Ｓ），Ｖ（Ｓ））を最大化する領域Ｓを出力するステップとをさらに含む請求項１記載のデータ間結合ルール導出方法。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】請求項１５
【補正方法】変更
【補正内容】
【請求項１５】切り出されるべき領域に含まれる最低限のデータ数である最小サポート数Ｕminを入力する手段

と、
前記切り出された領域Ｓに含まれるデータ数Ｕ（Ｓ）と前記最小サポート数Ｕminと比較する手段と、
前記比較の結果、Ｕmin≒Ｕ（Ｓ）であれば、当該領域Ｓを切り出されるべき領域として出力する手段と、
前記比較の結果、Ｕmim＞Ｕ（Ｓ）又はＵmim＜Ｕ（Ｓ）の場合には、新たな条件θ₁にて、前記領域切出手段が動作するように命ずる手段とを含む請求項１０記載のデータ間結合ルール導出装置。

＊【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】請求項１７
【補正方法】変更
【補正内容】
【請求項１７】切り出された前記領域Ｓに対し、
【数１２】

$$f(U(S),V(S)) = -\frac{V(S)}{Usum}\log\frac{V(S)}{Usum} - \frac{U(S)-V(S)}{Usum}\log\frac{U(S)-V(S)}{Usum} - \frac{Vsum-V(S)}{Usum-U(S)}\log\frac{Vsum-V(S)}{Usum-U(S)} - \frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}\log\frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}$$

（Ｕsumは前記平面全体のデータ数、Ｖsumは前記平面全体に含まれる前記真偽をとる属性が真のデータの数）を計算し、その値を前記領域Ｓに対応して記憶するエントロピ計算手段と、
変更された条件θにて前記領域切出手段及び前記エントロピ計算手段が動作するように命ずる手段と、
前記エントロピ計算手段に記憶されたｆ（Ｕ（Ｓ），Ｖ（Ｓ））を最大化する領域Ｓを出力する手段とを含む請求項１０記載のデータ間結合ルール導出装置。

【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】請求項２１
【補正方法】変更
【補正内容】
【請求項２１】コンピュータに、前記切り出された領域Ｓ内の各バケットのｖ(i,j)，ｕ(i,j)が、前記平面全体※

※のデータ数に対する前記平面全体の前記真偽をとる属性のデータ数の割合に等しくなるようｖ(i,j)を変更させるプログラムコード手段と、
当該変更されたｖ(i,j)及び入力された条件θ₁を用いて、コンピュータと前記領域切出プログラムコード手段が動作するように命じるプログラムコード手段とを有する請求項１９記載の記憶装置。

【手続補正５】
【補正対象書類名】明細書
【補正対象項目名】００２３
【補正方法】変更
【補正内容】
【００２３】ここで、切り出された前記領域Ｓに対し、
【数２３】

$$f(U(S),V(S)) = -\frac{V(S)}{Usum}\log\frac{V(S)}{Usum} - \frac{U(S)-V(S)}{Usum}\log\frac{U(S)-V(S)}{Usum} - \frac{Vsum-V(S)}{Usum-U(S)}\log\frac{Vsum-V(S)}{Usum-U(S)} - \frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}\log\frac{Usum-Vsum-U(S)+V(S)}{Usum-U(S)}$$

（Ｕsumは平面全体のデータ数、Ｖsumは平面全体に含まれる真偽をとる属性が真のデータの数）を計算し、その値を領域Ｓに対応して記憶するエントロピ計算ステップと、条件を変更して先の領域切出ステップとエントロピ計算ステップを実行するステップと、ｆ（Ｕ（Ｓ），Ｖ（Ｓ））を最大化する領域Ｓを出力するステップとを含むようにすれば、最適化エントロピ領域を見つけることができる。

【手続補正６】
【補正対象書類名】明細書
【補正対象項目名】００７４
【補正方法】変更
【補正内容】
【００７４】Ｄ．最適化エントロピ・ルールの場合
最適化エントロピ・ルールとは、領域の内部と外部との分割を考えた時、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである。よって、切り出された領域と平面全体のエントロピのゲイン（以下の

式）が最大となる領域を発見すればよい。

【数４２】

$$f(x,y) = -\frac{y}{a}\log\frac{y}{a} - \frac{x-y}{a}\log\frac{x-y}{a} - \frac{b-y}{a-x}\log\frac{b-y}{a-x} - \frac{a-b-x+y}{a-x}\log\frac{a-b-x+y}{a-x}$$

＊

＊このｘはＵ（Ｓ）、ｙはＶ（Ｓ）、ａはUsum、ｂはVsumである。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、θを変化させ、数４２を最大化するフォーカス・イメージを求めればよい。

フロントページの続き

(72)発明者　森本　康彦
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(72)発明者　森下　真一
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(72)発明者　徳山　豪
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内